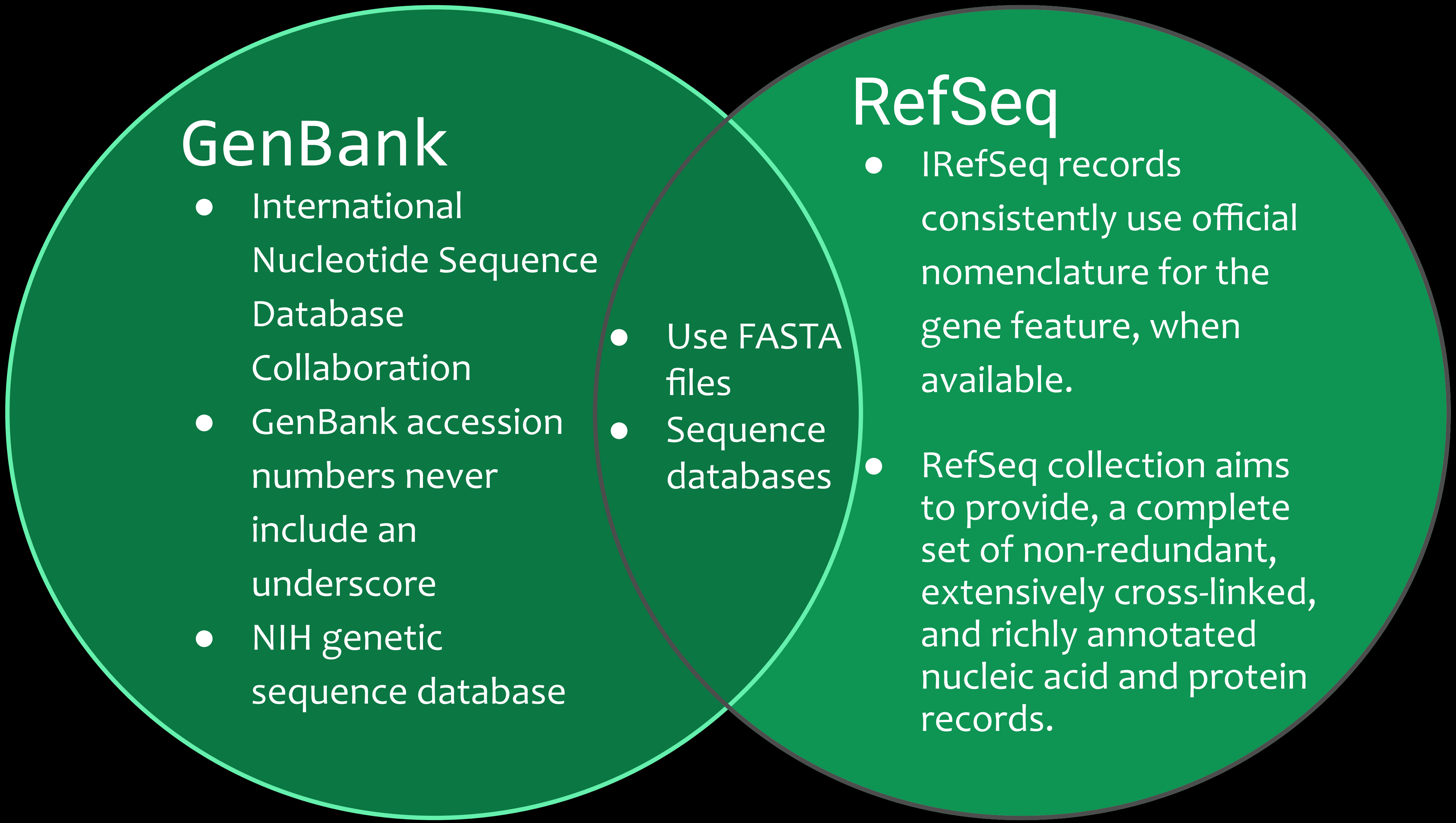
GenBank/RefSeQ Taxonomic Sequence Partitioning

Alexis Alvarez, Austin Torrence, Brandon Horner,
Chance Nelson, Tyler Malmon



GenBank

- International Nucleotide Sequence Database Collaboration
- GenBank accession numbers never include an underscore
- NIH genetic sequence database

- Use FASTA files
- Sequence databases

RefSeq

- IRefSeq records consistently use official nomenclature for the gene feature, when available.

- RefSeq collection aims to provide, a complete set of non-redundant, extensively cross-linked, and richly annotated nucleic acid and protein records.

GenBank/RefSeQ Taxonomic Sequence Partitioning

- Use of graph analysis
  - Common ancestry
  - Determining phylum
  - Common characteristics

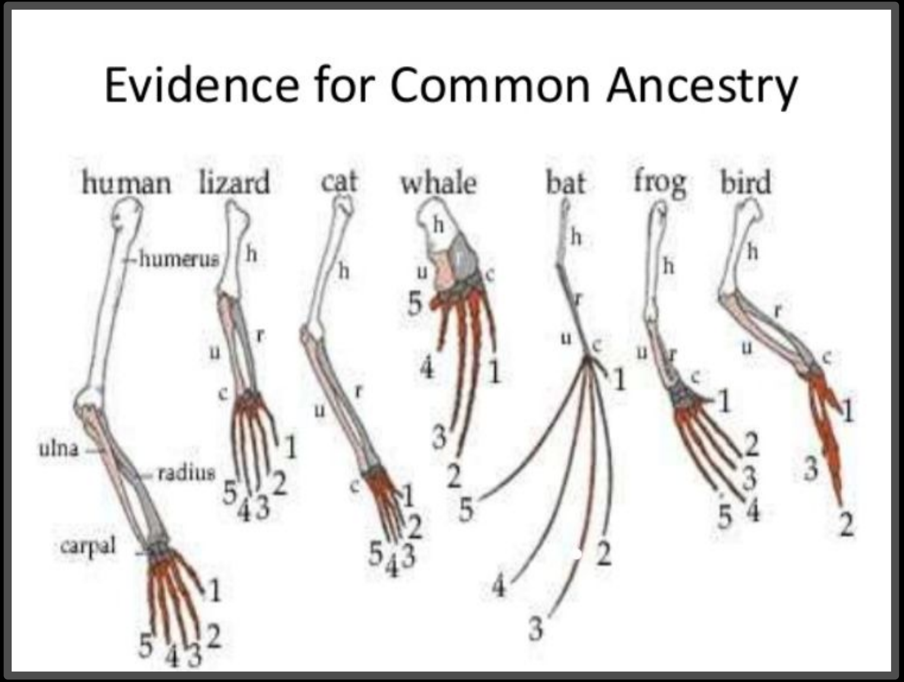
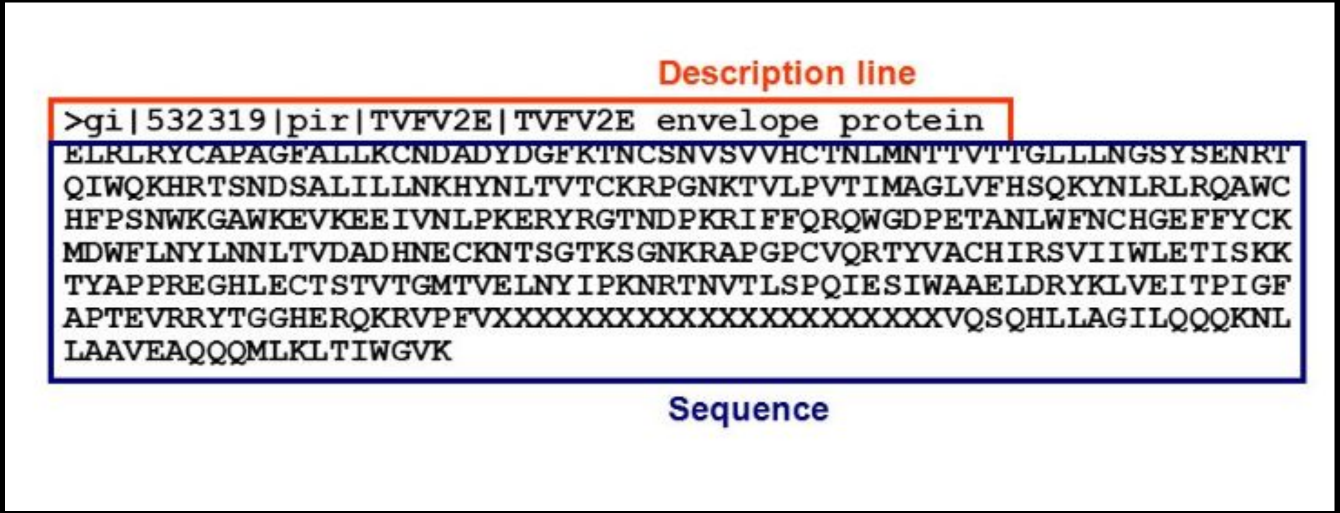


Figure 0: common ancestry

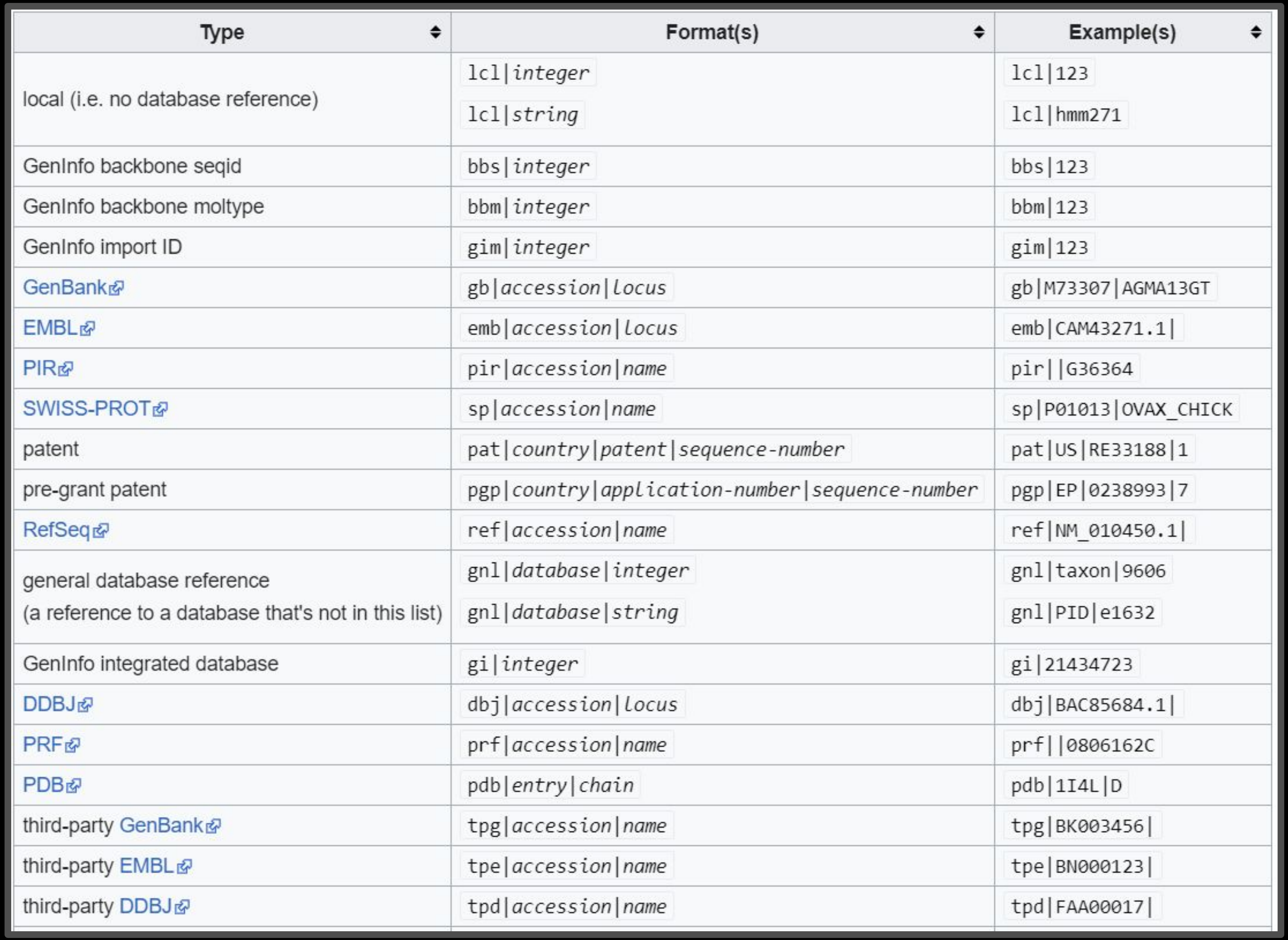
FASTA

- The universal format standard in the field of bioinformatics
- FASTA stores a variable number of sequence records


Description line

```
>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT
QIWQKHRTSNDSALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWGDPETANLWFNCHGEFFYCK
MDWFLNYLNNLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKK
TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLSPQIESIWAAELDRYKLVEITPIGF
APTEVRRYTGGHERQKRVPFVXXXXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL
LAAVEAQQQMLKLTIWGVK
```

Figure 1: FASTA Format



Type	Format(s)	Example(s)
local (i.e. no database reference)	lcl|*integer* lcl|*string*	lcl|123 lcl|hmm271
GenInfo backbone seqid	bbs|*integer*	bbs|123
GenInfo backbone moltype	bbm|*integer*	bbm|123
GenInfo import ID	gim|*integer*	gim|123
GenBank	gb|*accession*|*locus*	gb|M73307|AGMA13GT
EMBL	emb|*accession*|*locus*	emb|CAM43271.1|
PIR	pir|*accession*|*name*	pir||G36364
SWISS-PROT	sp|*accession*|*name*	sp|P01013|OVAX_CHICK
patent	pat|*country*|*patent*|*sequence-number*	pat|US|RE33188|1
pre-grant patent	pgp|*country*|*application-number*|*sequence-number*	pgp|EP|0238993|7
RefSeq	ref|*accession*|*name*	ref|NM_010450.1|
general database reference (a reference to a database that's not in this list)	gnl|*database*|*integer* gnl|*database*|*string*	gnl|taxon|9606 gnl|PID|e1632
GenInfo integrated database	gi|*integer*	gi|21434723
DDBJ	dbj|*accession*|*locus*	dbj|BAC85684.1|
PRF	prf|*accession*|*name*	prf||0806162C
PDB	pdb|*entry*|*chain*	pdb|1I4L|D
third-party GenBank	tpg|*accession*|*name*	tpg|BK003456|
third-party EMBL	tpe|*accession*|*name*	tpe|BN000123|
third-party DDBJ	tpd|*accession*|*name*	tpd|FAA00017|

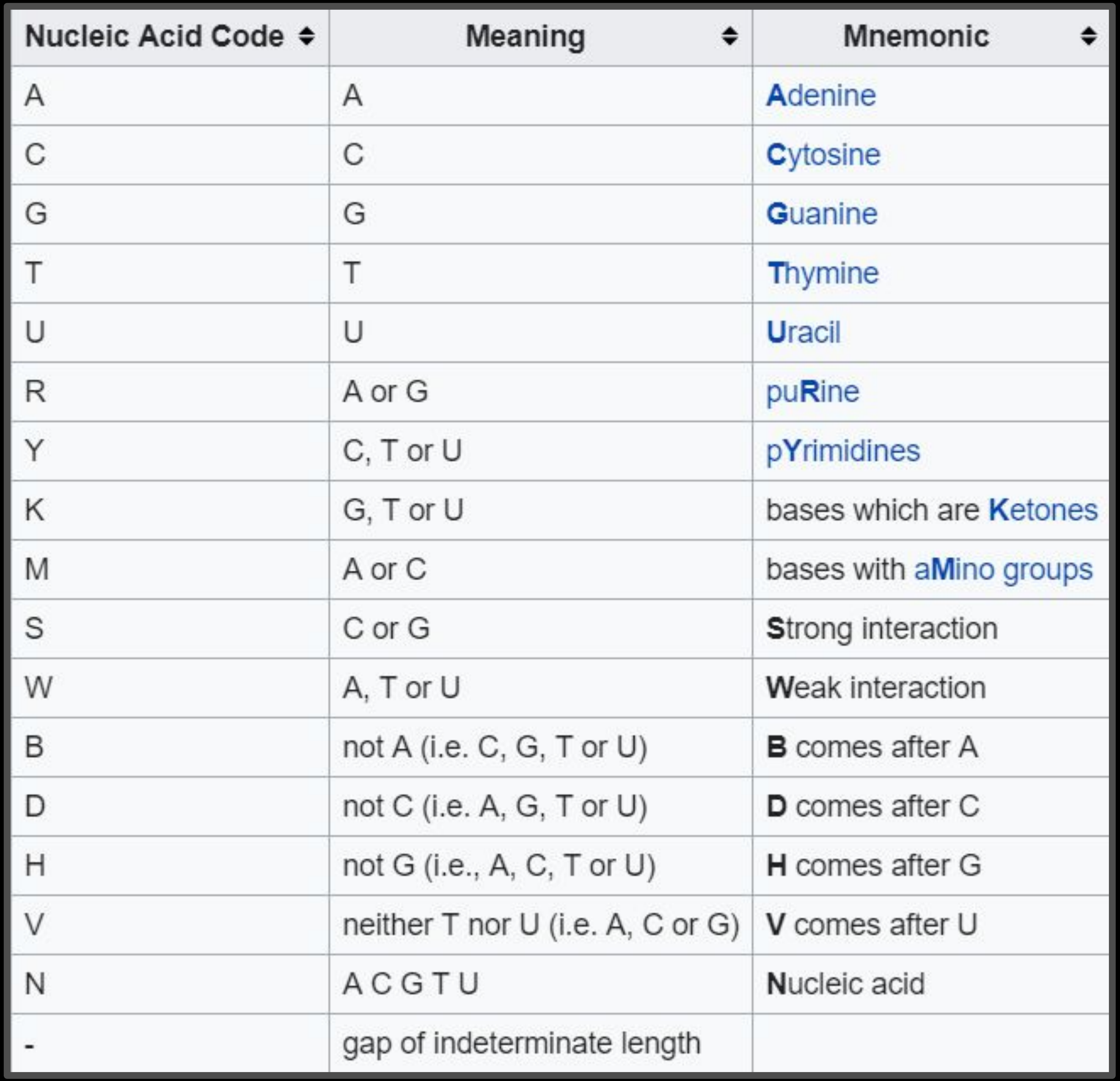
Figure 3: database identifiers table

NCBI Identifiers

NCBI FASTA defined format for sequence identifiers

Sequence Representation

Sequences may be protein sequences or nucleic acid sequences, and they can contain gaps or alignment characters



Nucleic Acid Code	Meaning	Mnemonic
A	A	**A**denine
C	C	**C**ytosine
G	G	**G**uanine
T	T	**T**hymine
U	U	**U**racil
R	A or G	pu**R**ine
Y	C, T or U	p**Y**rimidines
K	G, T or U	bases which are **K**etones
M	A or C	bases with a**M**ino groups
S	C or G	**S**trong interaction
W	A, T or U	**W**eak interaction
B	not A (i.e. C, G, T or U)	**B** comes after A
D	not C (i.e. A, G, T or U)	**D** comes after C
H	not G (i.e., A, C, T or U)	**H** comes after G
V	neither T nor U (i.e. A, C or G)	**V** comes after U
N	A C G T U	**N**ucleic acid
-	gap of indeterminate length	

Figure 4: Sequences table

Taxonomy

How Are Organisms Classified?

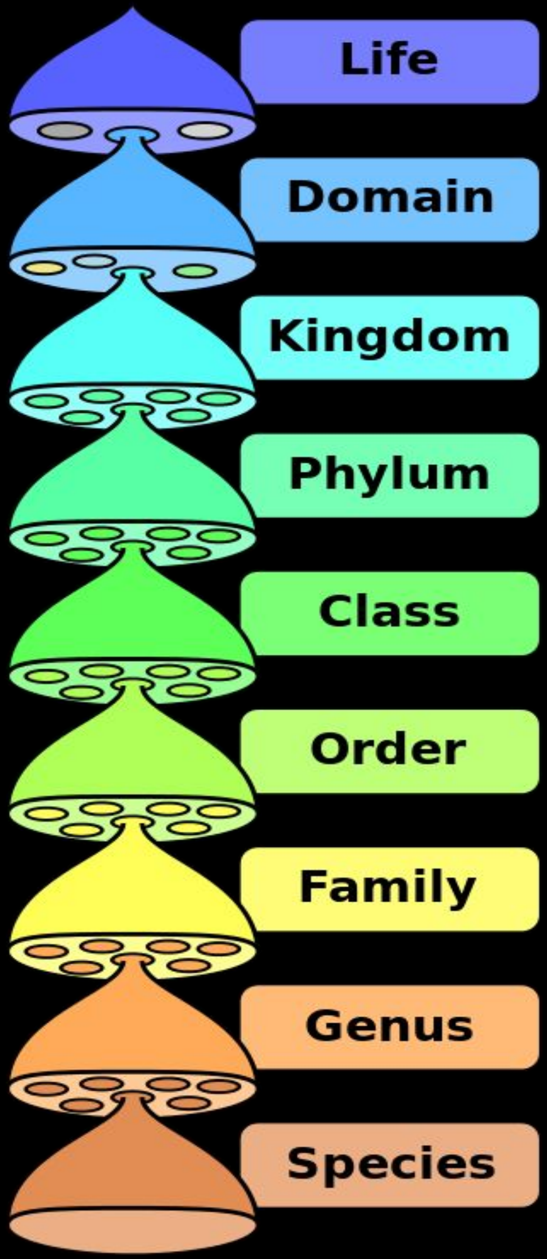
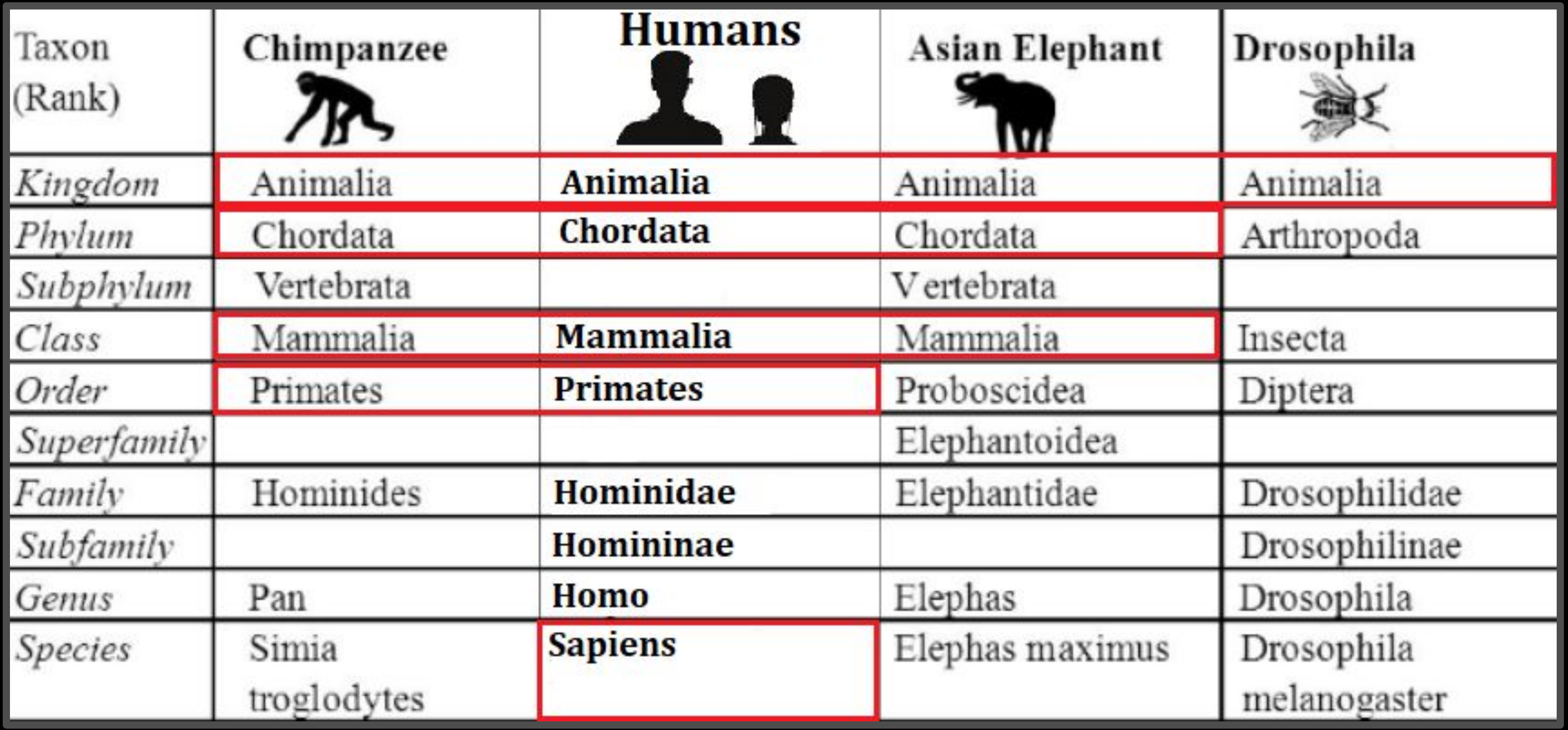


Figure 5: Modern Taxonomic Hierarchy

What are taxa?



Taxon (Rank)	Chimpanzee	Humans	Asian Elephant	Drosophila
Kingdom	Animalia	**Animalia**	Animalia	Animalia
Phylum	Chordata	**Chordata**	Chordata	Arthropoda
Subphylum	Vertebrata		Vertebrata	
Class	Mammalia	**Mammalia**	Mammalia	Insecta
Order	Primates	**Primates**	Proboscidea	Diptera
Superfamily			Elephantoidea	
Family	Hominides	**Hominidae**	Elephantidae	Drosophilidae
Subfamily		**Homininae**		Drosophilinae
Genus	Pan	**Homo**	Elephas	Drosophila
Species	Simia troglodytes	**Sapiens**	Elephas maximus	Drosophila melanogaster

Figure 5: Taxonomy table

Taxa

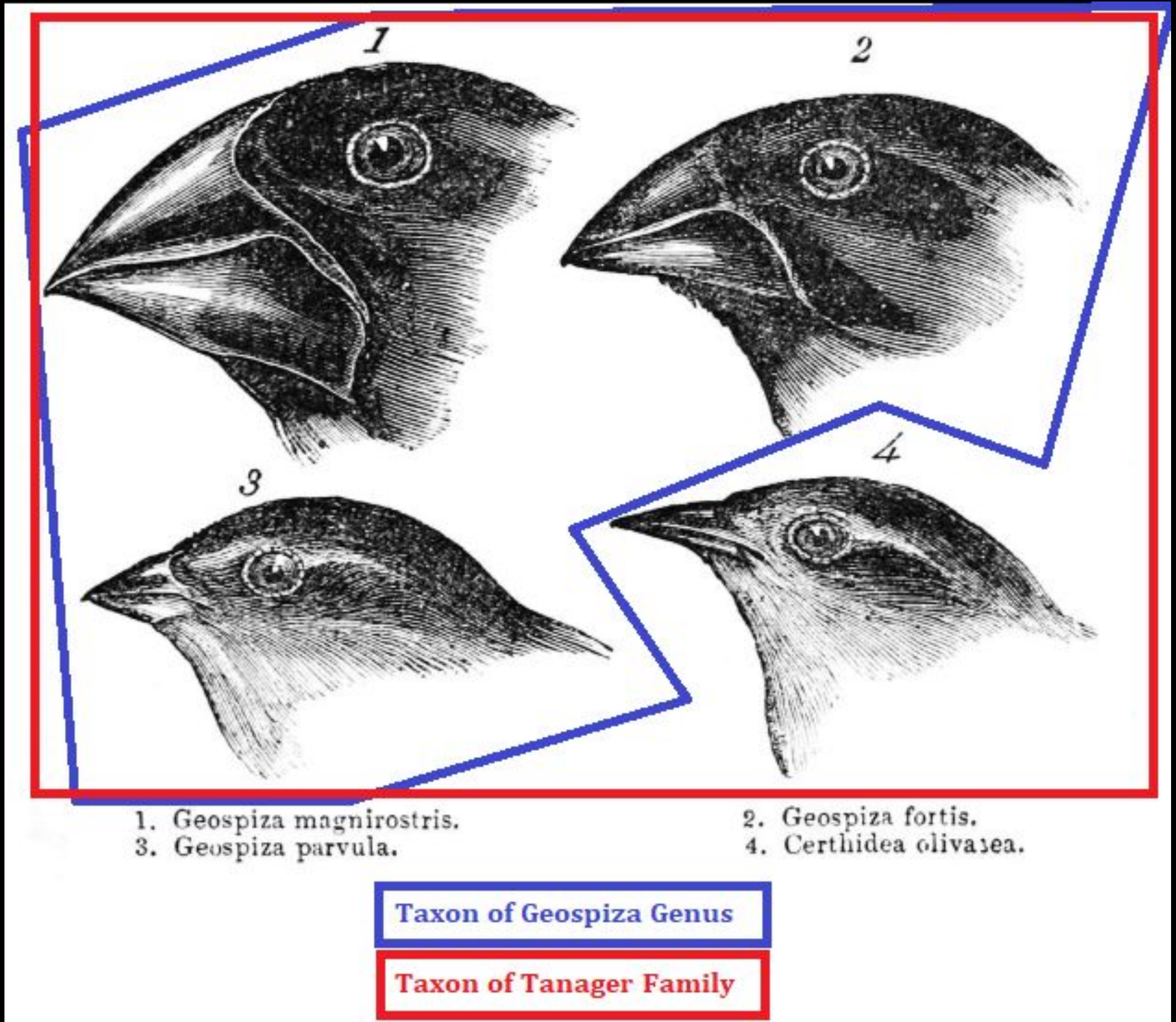
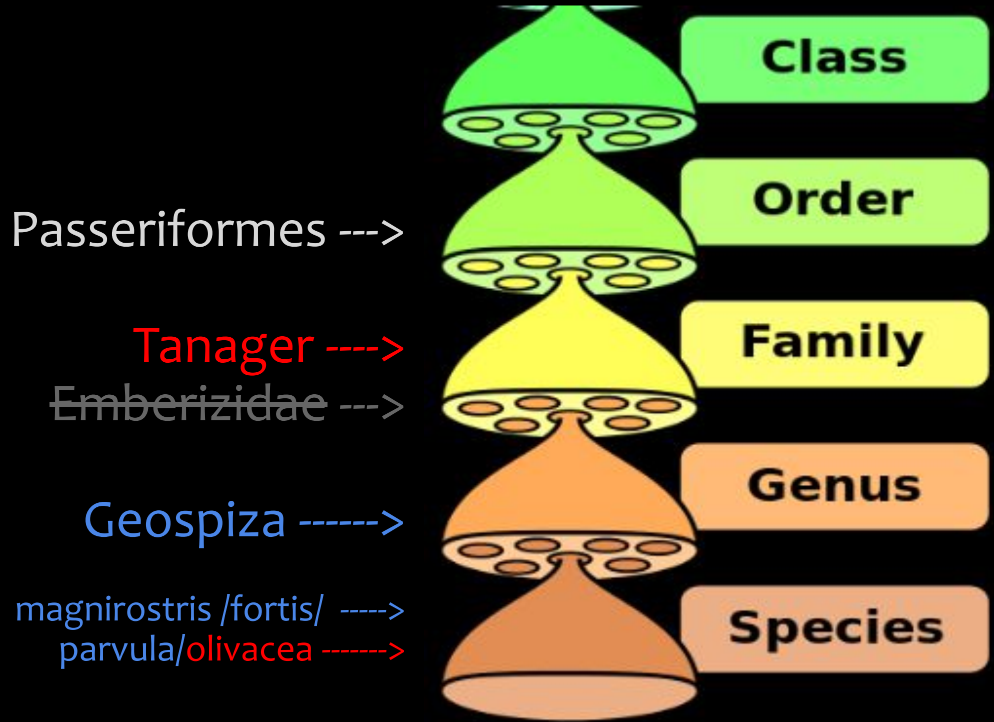


Figure 6: Darwin's finches by John Gould



Passeriformes --->

Tanager ---->

~~Emberizidae~~ --->

Geospiza ------>

magnirostris /fortis/ ----->

parvula/olivacea ------->

Taxonomic Sequence

- Graph complex networks of organisms
  - Nodes are organisms
  - Links are similarities between organisms
- Taxonomic sequences are better recognized as a phylogenetic trees.

Trees

- Fundamental Data Structure
  - Root Node
- Directed Graph
  - Single parent
  - Arbitrary amount of children

- Good for Hierarchical Organization

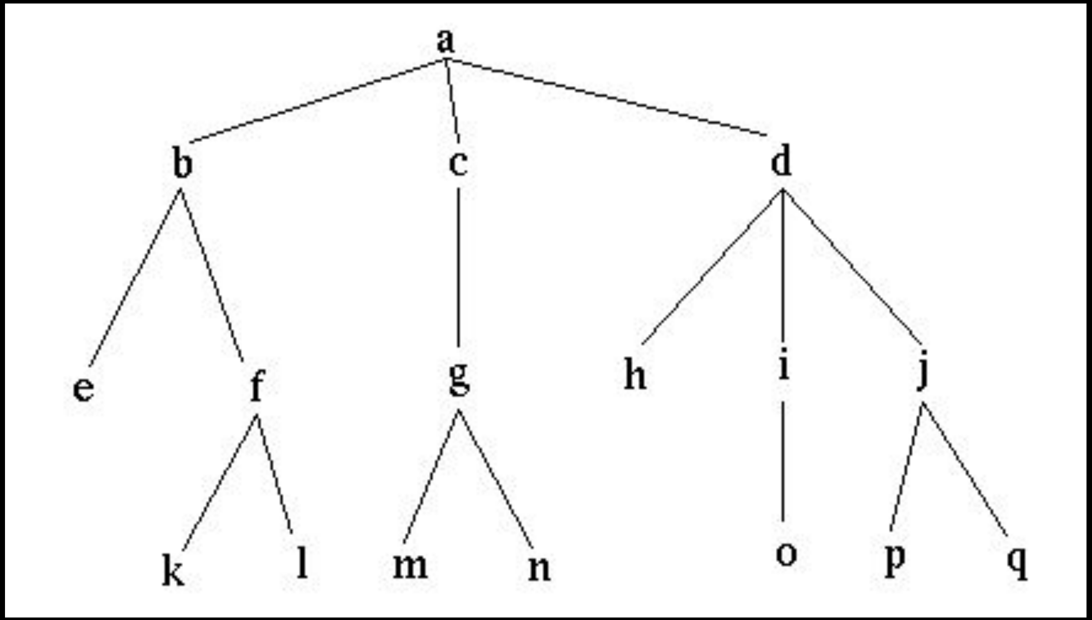


Figure 2: Example Tree

Phylogenetic trees

Phylogenetic trees are usually based on morphological or genetic homology

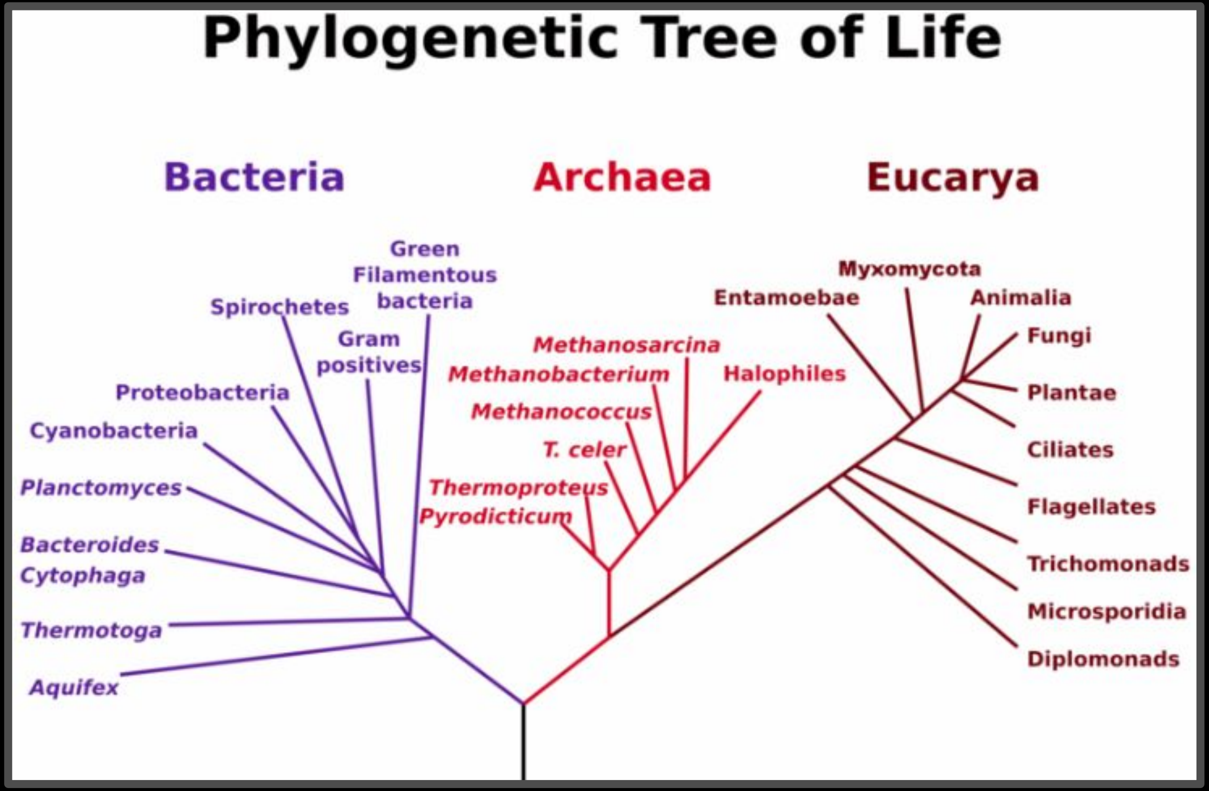


Figure 7: Phylogenetic tree of life

Development

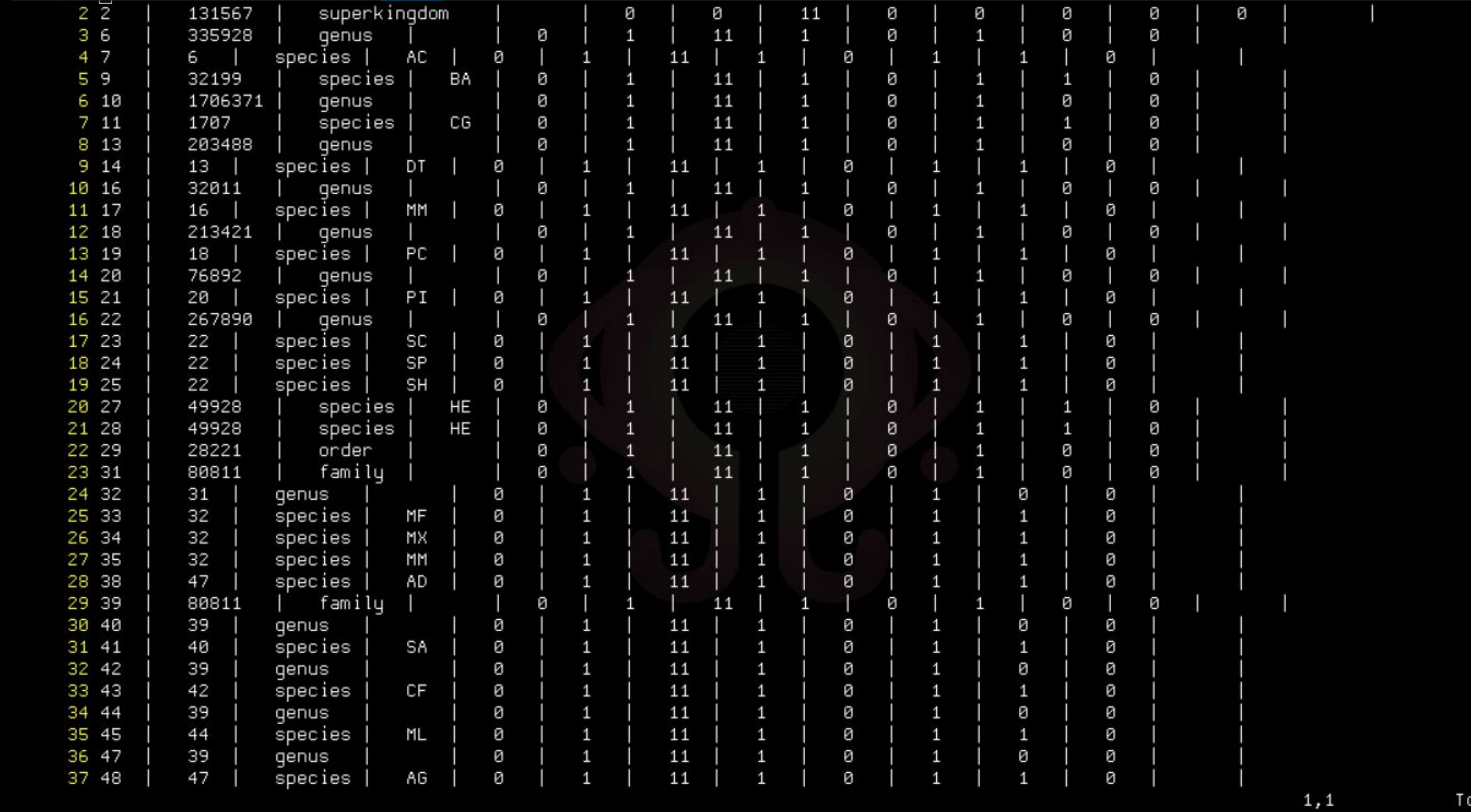
Plan and Current Status

‘GenBank/RefSeQ Taxonomic Sequence Partitioning’ - ???

- Retrieve NCBI Taxonomy Data

- Parse Taxa into Tree

- Make observations (...profit!)



```
2       |       131567  |       superkingdom    |               |       0       |       0       |       11      |       0       |       0       |       0       |       0       |       0       |               |
6       |       335928  |       genus   |               |       0       |       1       |       11      |       1       |       0       |       1       |       0       |       0       |               |
7       |       6       |       species |       AC      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
9       |       32199   |       species |       BA      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
10      |       1706371 |       genus   |               |       0       |       1       |       11      |       1       |       0       |       1       |       0       |       0       |               |
11      |       1707    |       species |       CG      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
13      |       203488  |       genus   |               |       0       |       1       |       11      |       1       |       0       |       1       |       0       |       0       |               |
14      |       13      |       species |       DT      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
16      |       32011   |       genus   |               |       0       |       1       |       11      |       1       |       0       |       1       |       0       |       0       |               |
17      |       16      |       species |       MM      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
18      |       213421  |       genus   |               |       0       |       1       |       11      |       1       |       0       |       1       |       0       |       0       |               |
19      |       18      |       species |       PC      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
20      |       76892   |       genus   |               |       0       |       1       |       11      |       1       |       0       |       1       |       0       |       0       |               |
21      |       20      |       species |       PI      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
22      |       267890  |       genus   |               |       0       |       1       |       11      |       1       |       0       |       1       |       0       |       0       |               |
23      |       22      |       species |       SC      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
24      |       22      |       species |       SP      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
25      |       22      |       species |       SH      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
27      |       49928   |       species |       HE      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
28      |       49928   |       species |       HE      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
29      |       28221   |       order   |               |       0       |       1       |       11      |       1       |       0       |       1       |       0       |       0       |               |
31      |       80811   |       family  |               |       0       |       1       |       11      |       1       |       0       |       1       |       0       |       0       |               |
32      |       31      |       genus   |               |       0       |       1       |       11      |       1       |       0       |       1       |       0       |       0       |               |
33      |       32      |       species |       MF      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
34      |       32      |       species |       MX      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
35      |       32      |       species |       MM      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
38      |       47      |       species |       AD      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
39      |       80811   |       family  |               |       0       |       1       |       11      |       1       |       0       |       1       |       0       |       0       |               |
40      |       39      |       genus   |               |       0       |       1       |       11      |       1       |       0       |       1       |       0       |       0       |               |
41      |       40      |       species |       SA      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
42      |       39      |       genus   |               |       0       |       1       |       11      |       1       |       0       |       1       |       0       |       0       |               |
43      |       42      |       species |       CF      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
44      |       39      |       genus   |               |       0       |       1       |       11      |       1       |       0       |       1       |       0       |       0       |               |
45      |       44      |       species |       ML      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
47      |       39      |       genus   |               |       0       |       1       |       11      |       1       |       0       |       1       |       0       |       0       |               |
48      |       47      |       species |       AG      |       0       |       1       |       11      |       1       |       0       |       1       |       1       |       0       |               |
                                                                                                                                    1,1           To
```

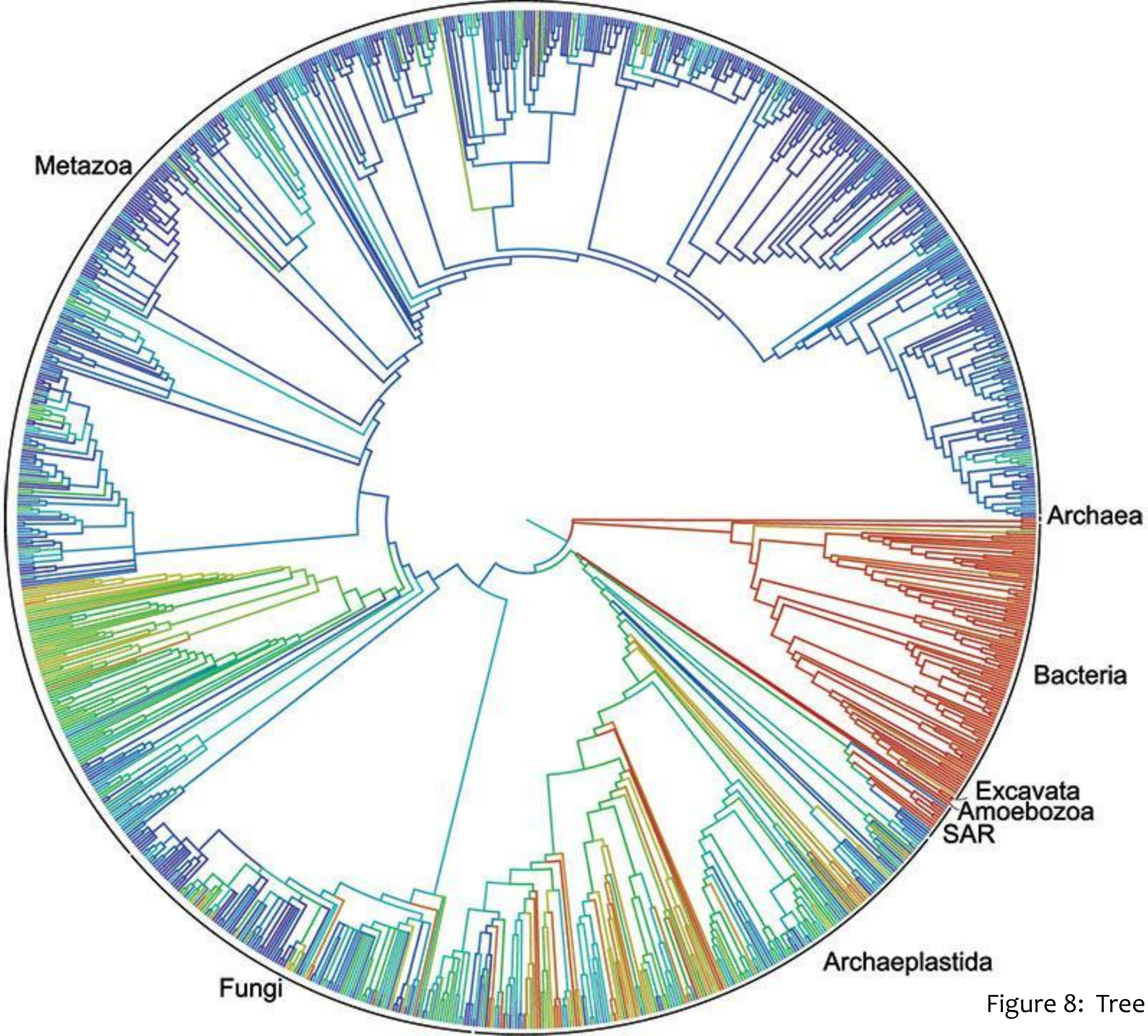
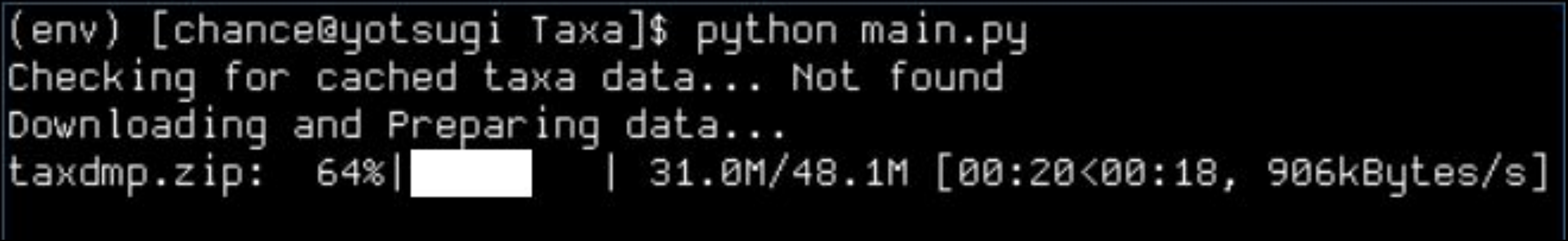


Figure 8: Tree of Life

Retrieve NCBI Data

- URLLIB + TQDM

- Caching
  - MD5

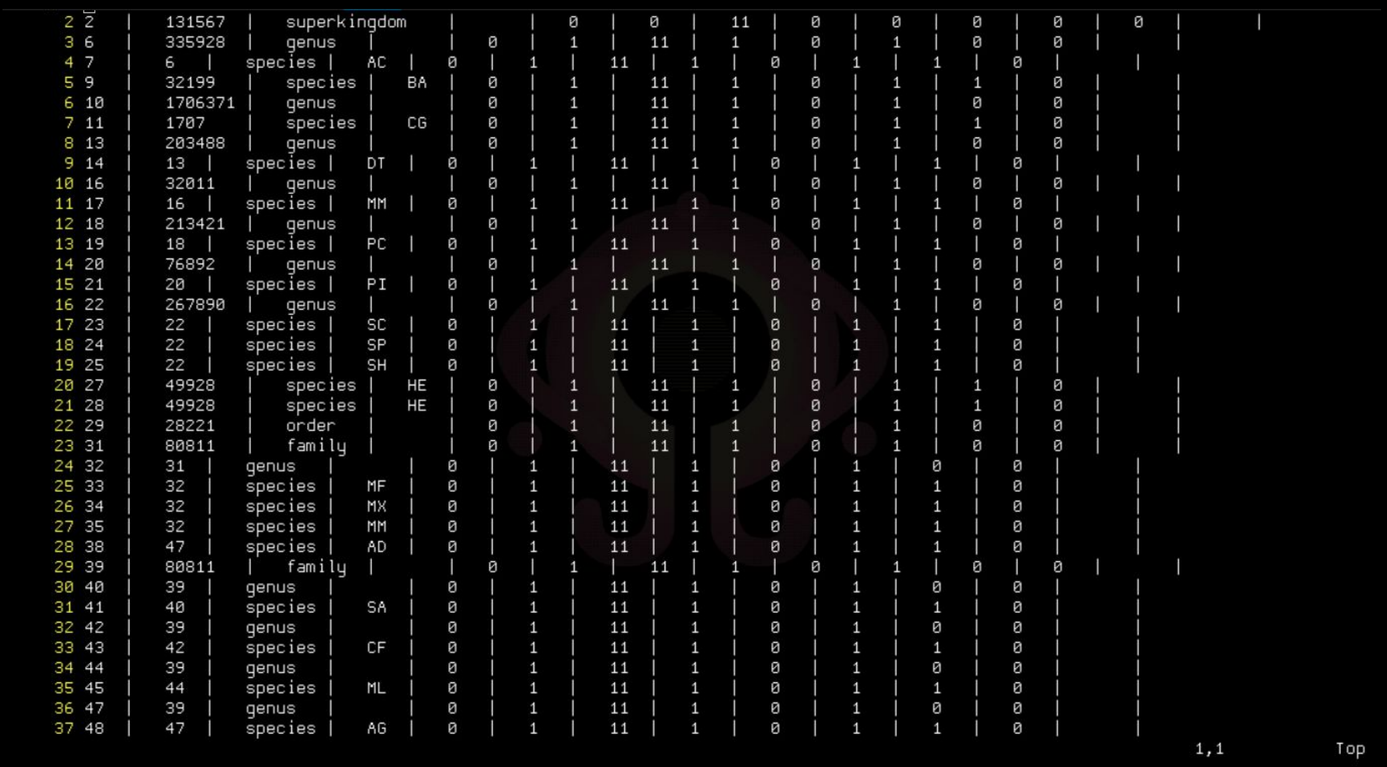


```
(env) [chance@yotsugi Taxa]$ python main.py
Checking for cached taxa data... Not found
Downloading and Preparing data...
taxdmp.zip:  64%|█████     | 31.0M/48.1M [00:20<00:18, 906kBytes/s]
```

Figure 9: retrieving data

Parsing into Tree

- CSV Parsing
- NetworkX
- pyplot
    - slow



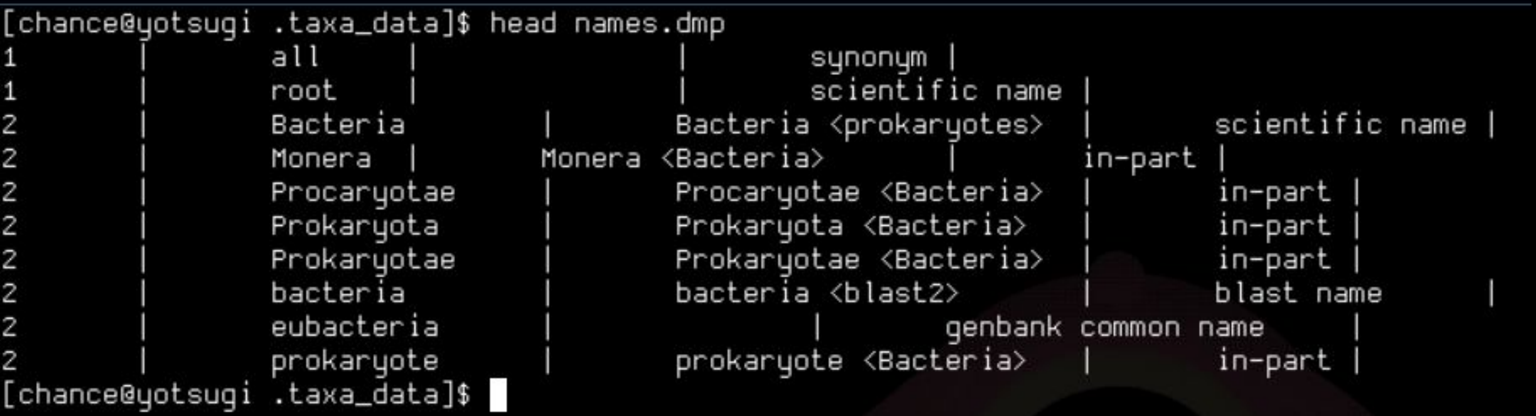
```
2   |   131567  |   superkingdom    |       |   0   |   0   |   11  |   0   |   0   |   0   |   0   |   0   |       |
6   |   335928  |   genus   |       |   0   |   1   |   11  |   1   |   0   |   1   |   0   |   0   |       |
7   |   6   |   species |   AC  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
9   |   32199   |   species |   BA  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
10  |   1706371 |   genus   |       |   0   |   1   |   11  |   1   |   0   |   1   |   0   |   0   |       |
11  |   1707    |   species |   CG  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
13  |   203488  |   genus   |       |   0   |   1   |   11  |   1   |   0   |   1   |   0   |   0   |       |
14  |   13  |   species |   DT  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
16  |   32011   |   genus   |       |   0   |   1   |   11  |   1   |   0   |   1   |   0   |   0   |       |
17  |   16  |   species |   MM  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
18  |   213421  |   genus   |       |   0   |   1   |   11  |   1   |   0   |   1   |   0   |   0   |       |
19  |   18  |   species |   PC  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
20  |   76892   |   genus   |       |   0   |   1   |   11  |   1   |   0   |   1   |   0   |   0   |       |
21  |   20  |   species |   PI  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
22  |   267890  |   genus   |       |   0   |   1   |   11  |   1   |   0   |   1   |   0   |   0   |       |
23  |   22  |   species |   SC  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
24  |   22  |   species |   SP  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
25  |   22  |   species |   SH  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
27  |   49928   |   species |   HE  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
28  |   49928   |   species |   HE  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
29  |   28221   |   order   |       |   0   |   1   |   11  |   1   |   0   |   1   |   0   |   0   |       |
31  |   80811   |   family  |       |   0   |   1   |   11  |   1   |   0   |   1   |   0   |   0   |       |
32  |   31  |   genus   |       |   0   |   1   |   11  |   1   |   0   |   1   |   0   |   0   |       |
33  |   32  |   species |   MF  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
34  |   32  |   species |   MX  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
35  |   32  |   species |   MM  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
38  |   47  |   species |   AD  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
39  |   80811   |   family  |       |   0   |   1   |   11  |   1   |   0   |   1   |   0   |   0   |       |
40  |   39  |   genus   |       |   0   |   1   |   11  |   1   |   0   |   1   |   0   |   0   |       |
41  |   40  |   species |   SA  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
42  |   39  |   genus   |       |   0   |   1   |   11  |   1   |   0   |   1   |   0   |   0   |       |
43  |   42  |   species |   CF  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
44  |   39  |   genus   |       |   0   |   1   |   11  |   1   |   0   |   1   |   0   |   0   |       |
45  |   44  |   species |   ML  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
47  |   39  |   genus   |       |   0   |   1   |   11  |   1   |   0   |   1   |   0   |   0   |       |
48  |   47  |   species |   AG  |   0   |   1   |   11  |   1   |   0   |   1   |   1   |   0   |       |
                                                                                                1,1           Top
```

Figure 10: Taxonomy nodelist

Challenges and Plans

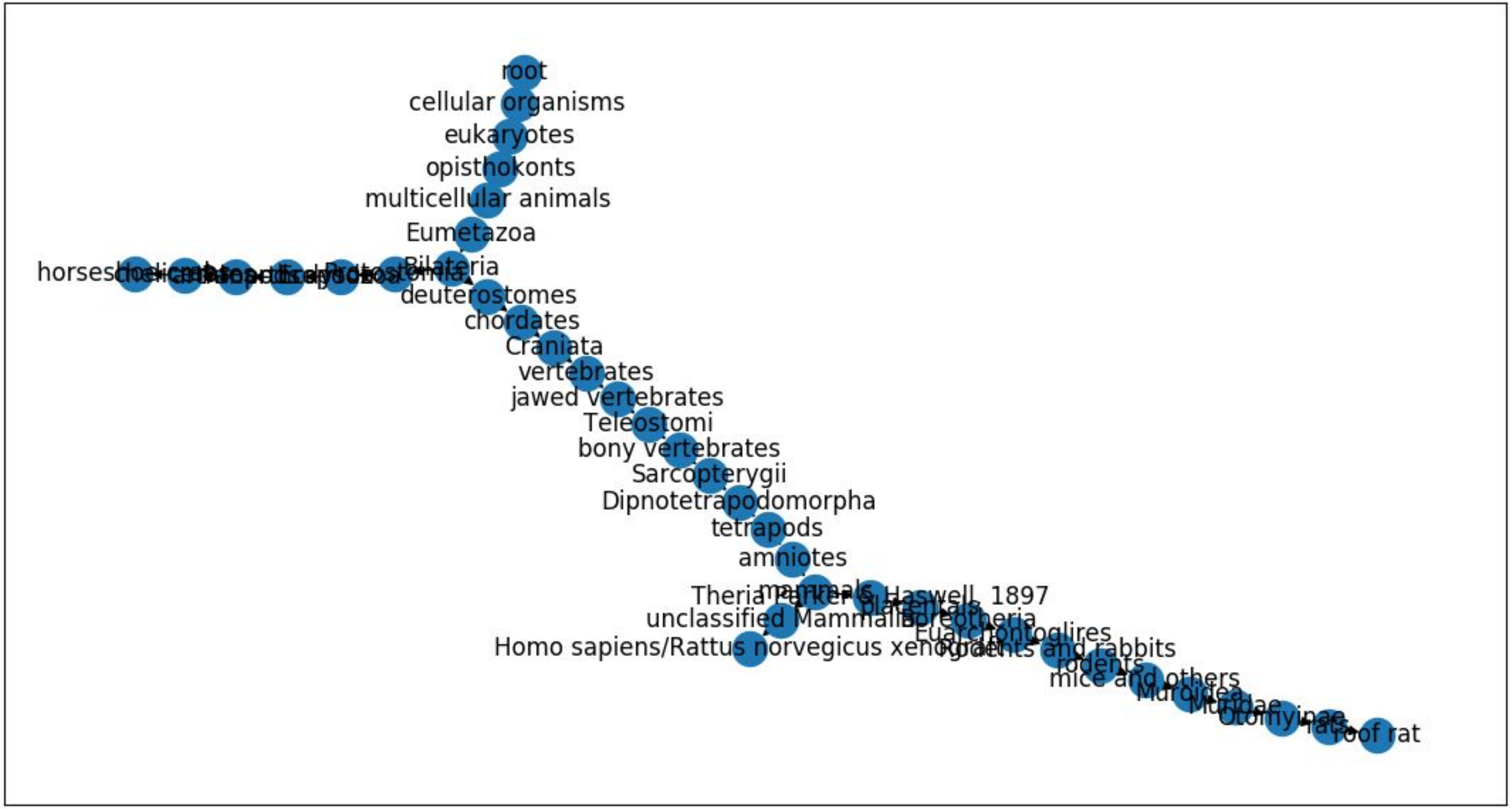
- Slow Plot Generation
  - More Power!
  - Gephi
  - Subtrees
- Taxa IDs
  - Name Associations
- Analysis
  - PageRank
  - Node Distributions
- QOL Improvements
  - Name lookup
  - Global tree caching

Figure 11: names.dmp



```
[chance@yotsugi .taxa_data]$ head names.dmp
1       |       all     |               |       synonym |
1       |       root    |               |       scientific name |
2       |       Bacteria        |       Bacteria <prokaryotes>  |       scientific name |
2       |       Monera  |       Monera <Bacteria>       |       in-part |
2       |       Procaryotae     |       Procaryotae <Bacteria>  |       in-part |
2       |       Prokaryota      |       Prokaryota <Bacteria>   |       in-part |
2       |       Prokaryotae     |       Prokaryotae <Bacteria>  |       in-part |
2       |       bacteria        |       bacteria <blast2>       |       blast name      |
2       |       eubacteria      |               |       genbank common name     |
2       |       prokaryote      |       prokaryote <Bacteria>   |       in-part |
[chance@yotsugi .taxa_data]$ █
```

Figure 12: Monsoon Logo


root
cellular organisms
eukaryotes
opisthokonts
multicellular animals
Eumetazoa
Bilateria
Protostomia
deuterostomes
chordates
Craniata
vertebrates
jawed vertebrates
Teleostomi
bony vertebrates
Sarcopterygii
Dipnotetrapodomorpha
tetrapods
amniotes
mammals
Theria Parker & Haswell, 1897
unclassified Mammalia
Homo sapiens/Rattus norvegicus xenograft
placentals
Boreotheria
Euarchontoglires
Rodents and rabbits
rodents
mice and others
Muroidea
Muridae
Otomyinae
rats
roof rat

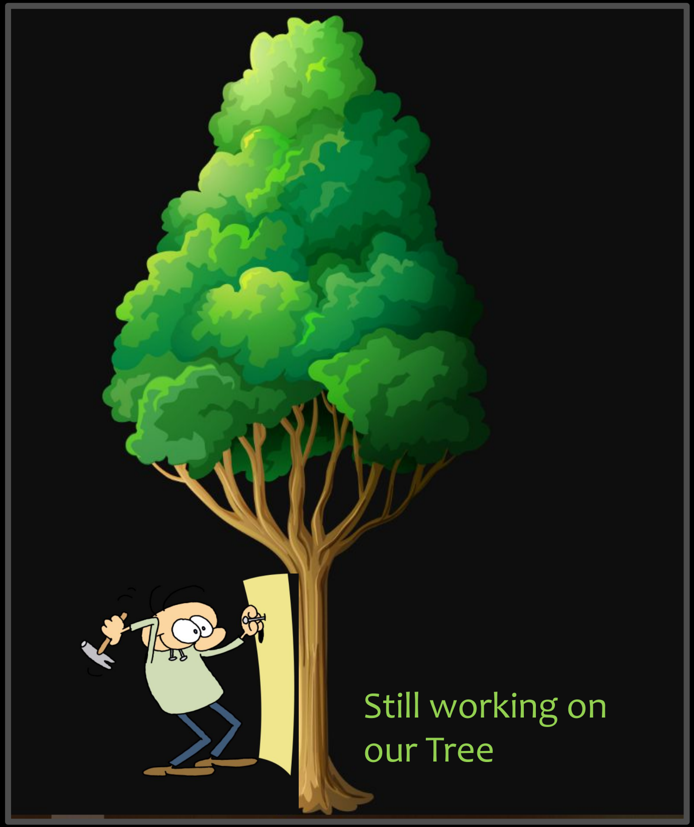
GitHub

https://github.com/bioinformatics-spr19/Taxa

In conclusion

- Taxonomy is a work in progress.

Thank You



Still working on our Tree