Scoring matrix –BLOSUM 62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Basic concept of motif identification 2.

How do we represent the position specific preference ?

```
BID_MOUSE    I A R H L A Q I G D E M
BAD_MOUSE    Y G R E L R R M S D E F
BAK_MOUSE    V G R Q L A L I G D D I
BAXB_HUMAN   L S E C L K R I G D E L
BimS         I A Q E L R R I G D E F
HRK_HUMAN    T A A R L K A L G D E L
Egl-1        I G S K L A A M C D D F
```

Statistical representation

G: 5 -> 71%

S: 1 -> 14 %

C: 1 -> 14 %

Identifying shared motifs using MEME -Multiple EM for Motif Elicitation

- Identifies statistically significant motif(s) in a set of sequences.
- Motifs shared by proteins.
  - Protein family.
  - Mediate interaction between different protein.
- Motifs shared by DNA sequences binding to certain transcription factor (ChIP-Seq).

Two search examples

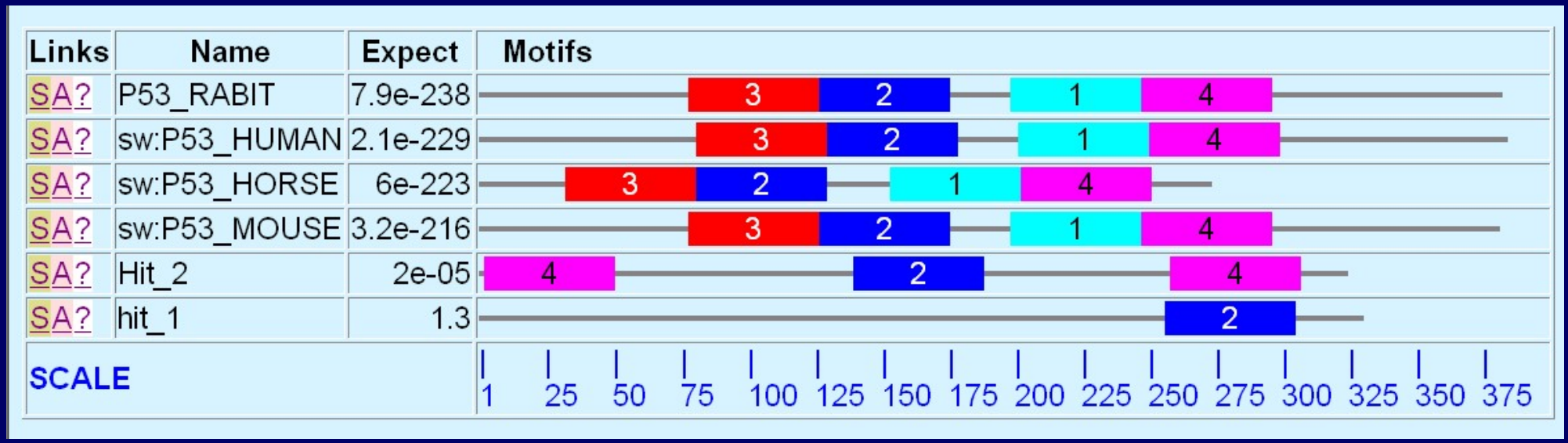
Set1: Mammalian P53 plus mosquito hits

Set2: Diverse set of P53 plus mosquito hits

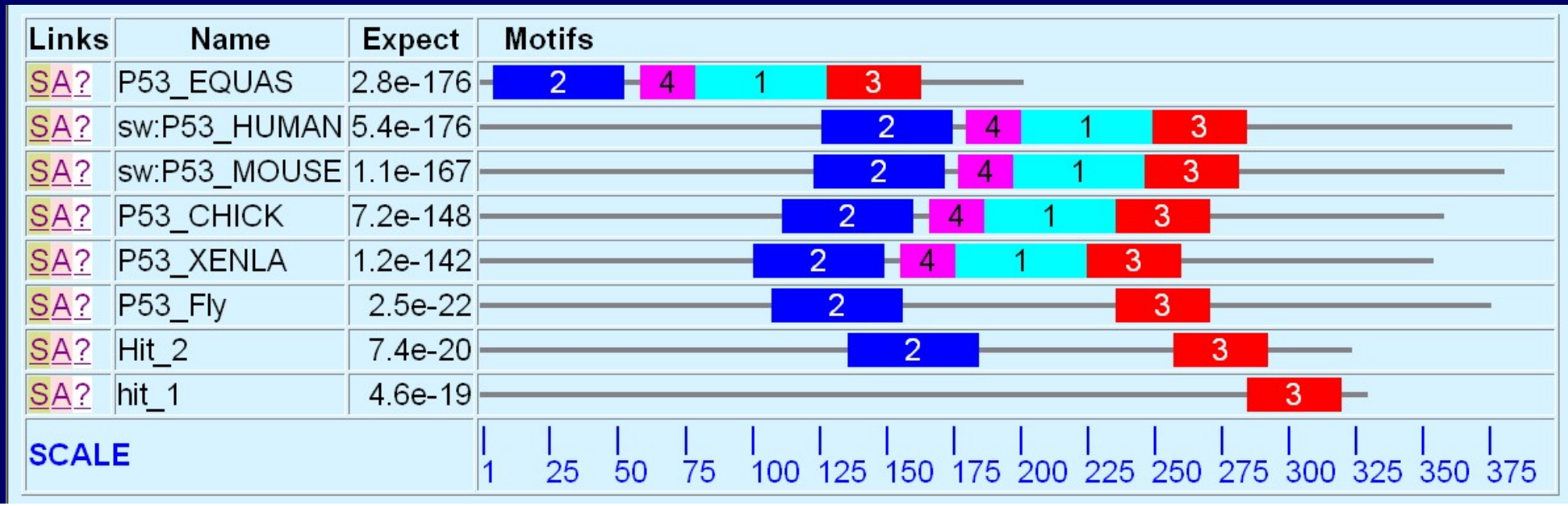
- The outcome of the search is dependent on the inputting set of sequences.
- Compose the inputting set based on your research needs.

Selection of sequences determines the model

Set1: Mammalian P53 plus mosquito hits



Set2: Diverse set of P53 plus mosquito hits



Building Phylogenetic Trees

What is a phylogenetic Tree?

- How the observed differences between sequences are developed through evolution.

- The distance between sequences.

Steps of Building Phylogenetic Trees

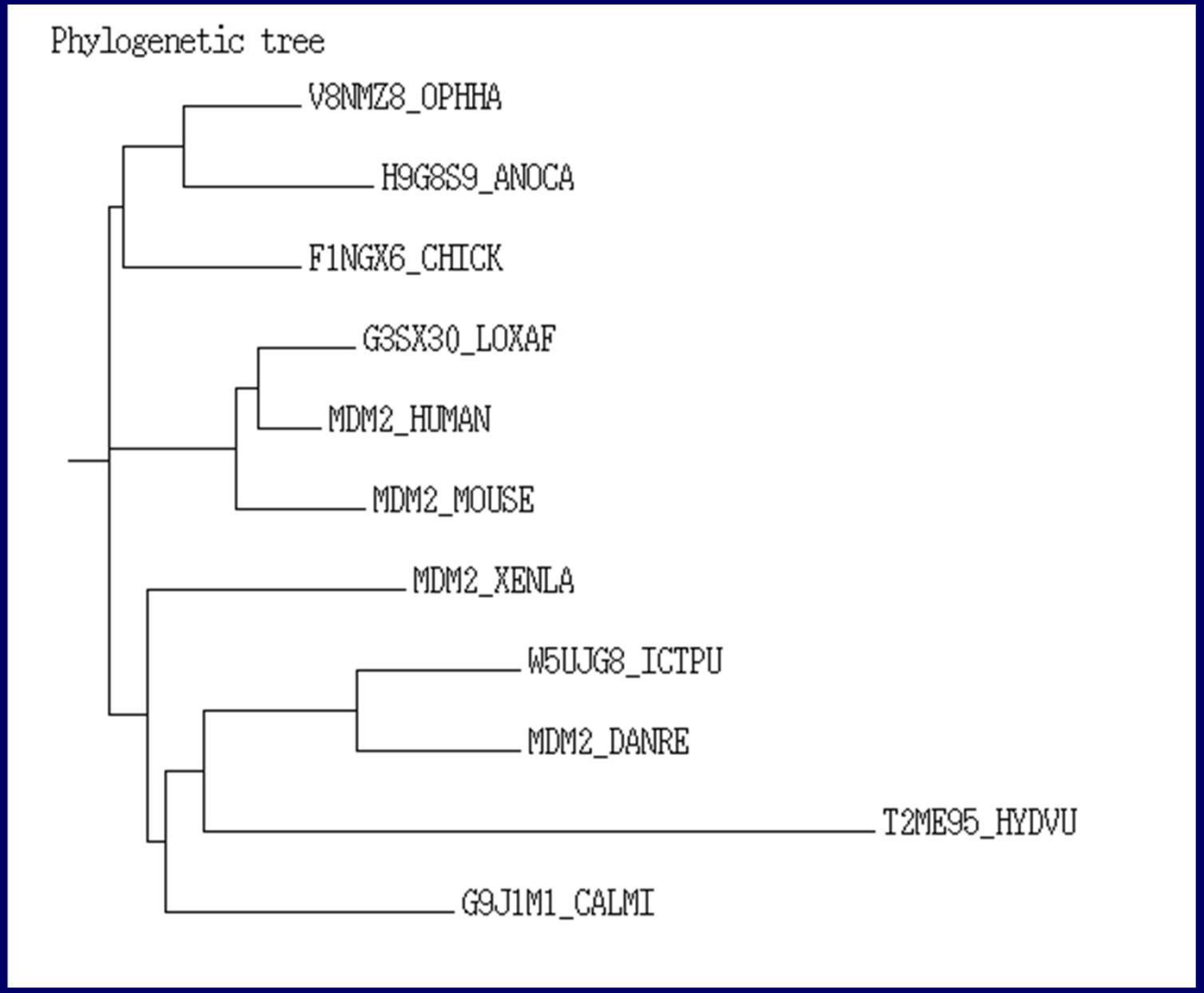
1. Collect sequences in one FASTA format file.
2. Perform multiple sequence alignment (global).
3. Draw phylogenetic trees (different approaches).
4. Bootstrapping the phylogenetic tree
5. View and edit the tree for presentation.

Building Phylogenetic Trees

Practice:

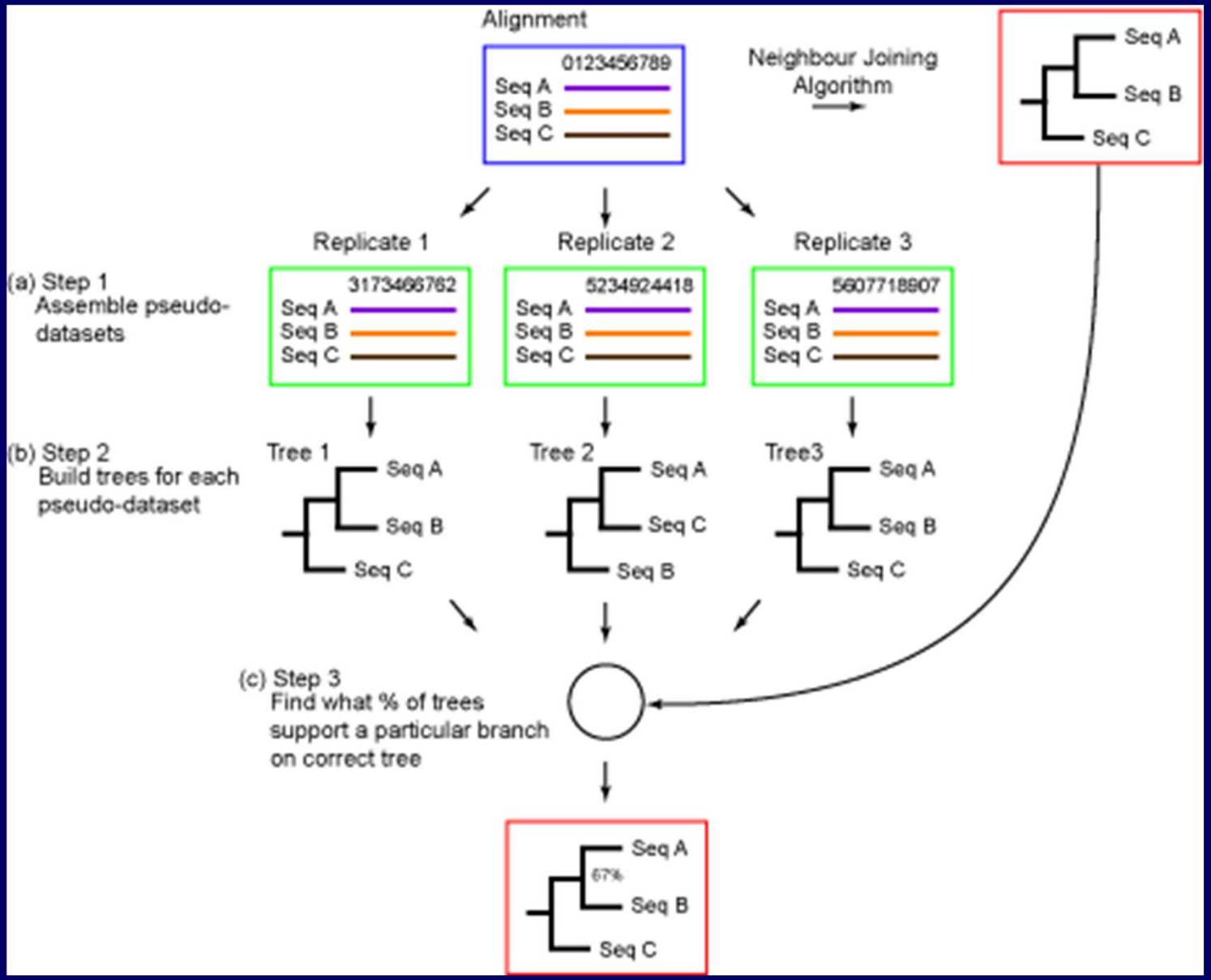
1. Load sequence in FASTA format to Clustal Omega to perform alignment.
2. Download the tree file save in your GMS6014/XXX folder.
3. View the tree with the Phylodendron tree printer.

Phylogenetic Trees



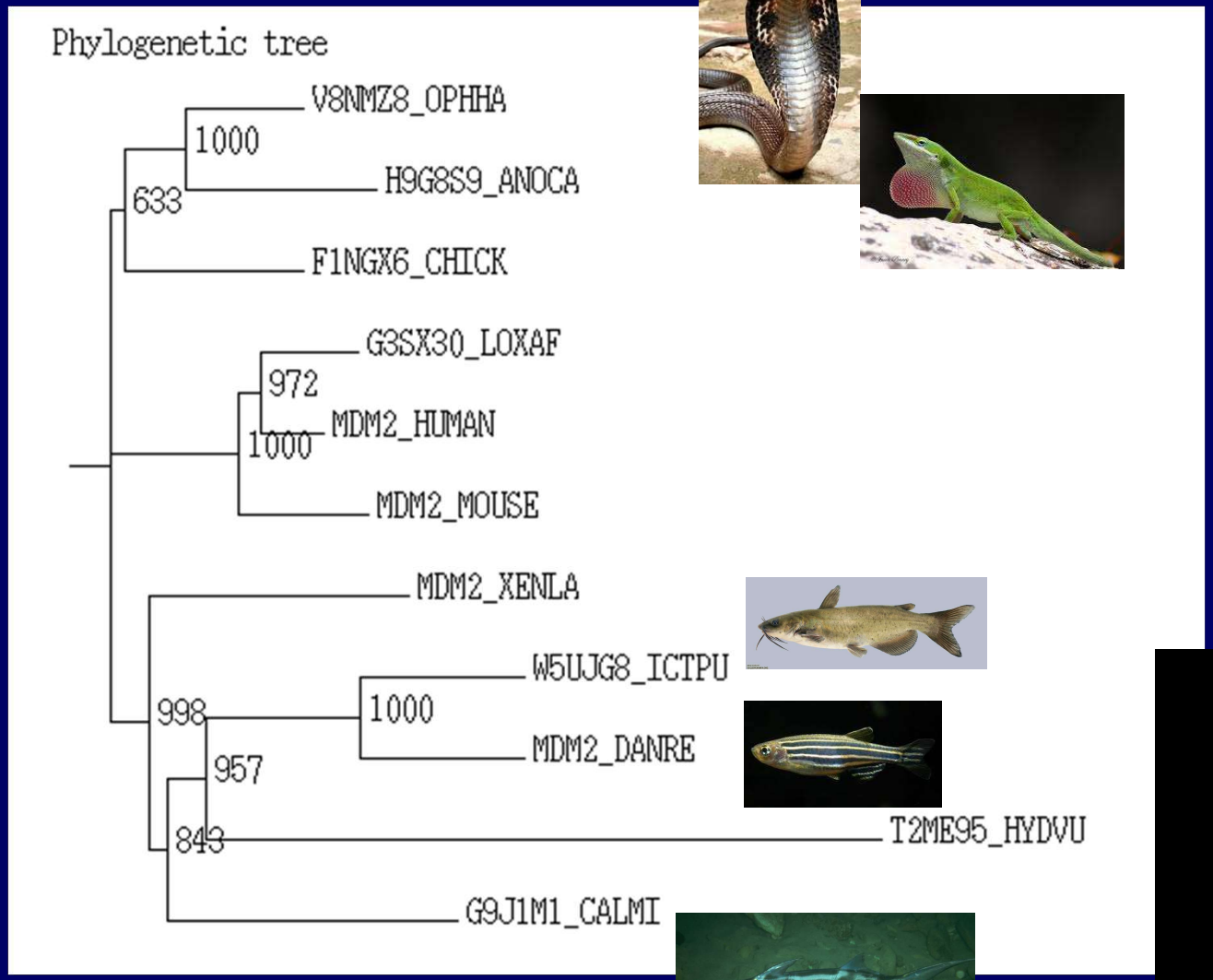
LOXAF: African Elephant

Bootstrapping

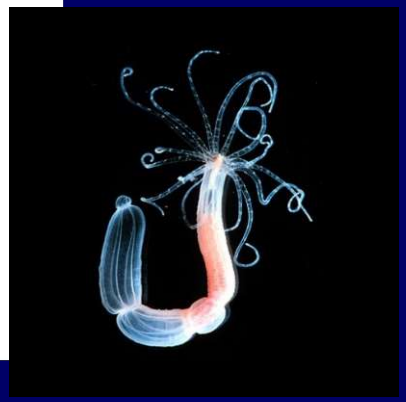


http://phylogenetictrees.com/

Phylogenetic Trees – boot strap



LOXAF: African Elephant



Building Phylogenetic Trees

Observe:

Extracting sequence, alignment, and tree building with Jalview.

https://www.jalview.org/

Practice –Download GEO dataset with job file

Retrieval of information.

- Using web interface.
- Using FTP client

- Using command line tools.
  - Generic Linux file transfer tools - always available in Linux/MacOs.
  - Specialized tool – fastq-dump
  - Let the script do the job – when you need large amount of files or large file that will take hours to download.

RNA-Seq – Download dataset

- Commend line (with a few samples):
  - $ module load sra
  - $ fastq-dump --gzip SRRxxxx SRRyyyy
- With the .sbatch job file (for large data set)
  - $sbatch myjob.sbatch
  - Use “$ squeue –u <yourUserName>” to monitor progress.
  - Use “$ls –l” to make sure files size are correct.

RNA-Seq – Download dataset

```
#!/bin/sh
#SBATCH --job-name=GetSRA
#SBATCH --mail-type=ALL
#SBATCH --mail-user=xxxxx@ufl.edu
#SBATCH --output=GetSRA_%j.log
#SBATCH -t 12:00:00
#SBATCH --cpus-per-task=1
#SBATCH --mem-per-cpu=3gb

pwd; date

module load sra/2.10.3

fastq-dump --gzip SRR1618640 SRR1618641 SRR1618642 SRR1618643
```

Transfer the file to your folder in HiPerGator and submit the job ($sbatch *filename*)

Practice –retrieve genome to HiPerGator with wget