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).**
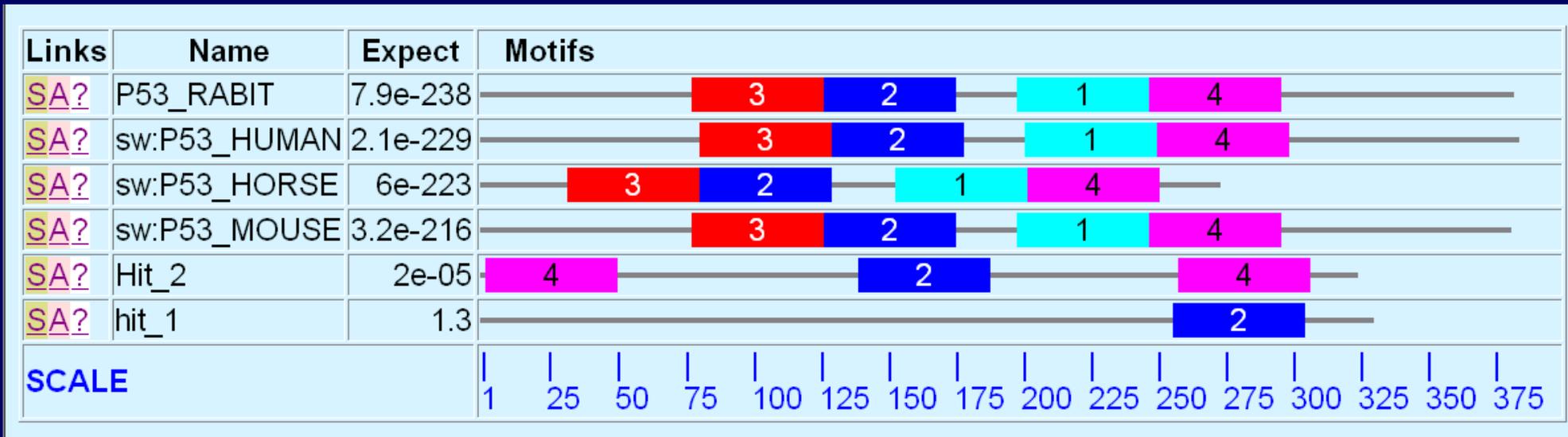
Two search examples

Set1: Mammalian P53 plus mosquito hits
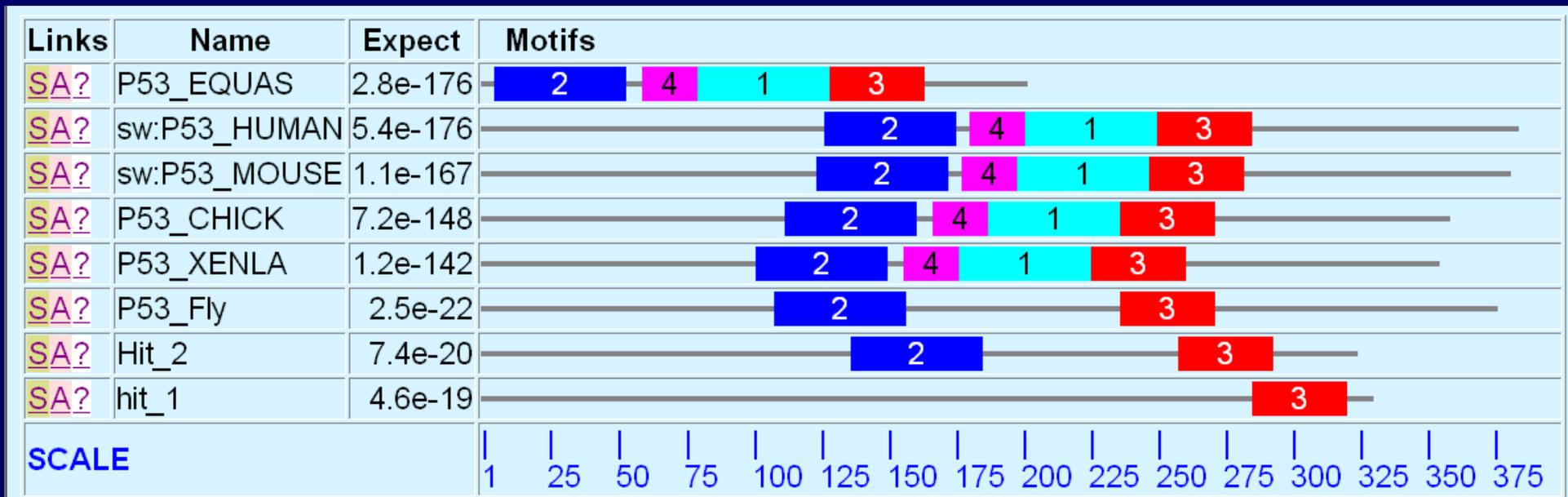
Set2: Diverse set of P53 plus mosquito hits

➤ 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

Links	Name	Expect	Motifs
SA?	P53_RABIT	7.9e-238	3 2 1 4
SA?	sw:P53_HUMAN	2.1e-229	3 2 1 4
SA?	sw:P53_HORSE	6e-223	3 2 1 4
SA?	sw:P53_MOUSE	3.2e-216	3 2 1 4
SA?	Hit_2	2e-05	4 2 4
SA?	hit_1	1.3	2
SCALE			1 25 50 75 100 125 150 175 200 225 250 275 300 325 350 375

Set2: Diverse set of P53 plus mosquito hits

Links	Name	Expect	Motifs
SA?	P53_EQUAS	2.8e-176	2 4 1 3
SA?	sw:P53_HUMAN	5.4e-176	2 4 1 3
SA?	sw:P53_MOUSE	1.1e-167	2 4 1 3
SA?	P53_CHICK	7.2e-148	2 4 1 3
SA?	P53_XENLA	1.2e-142	2 4 1 3
SA?	P53_Fly	2.5e-22	2 3
SA?	Hit_2	7.4e-20	2 3
SA?	hit_1	4.6e-19	3
SCALE			1 25 50 75 100 125 150 175 200 225 250 275 300 325 350 375

Building Phylogenetic Trees

What is a phylogenetic Tree?

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

➢ Focus on the distance (difference) between sequences.

➢ Select a representative group of 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.

Steps of Multiple Sequence Alignment by Clustal O

1. **Estimate pairwise distances** between sequences (how dissimilar they are) to build a distance matrix.

2. Build a **guide tree** from those distances (typically a fast clustering/tree method).

3. **Progressively align** starting from the closest sequences, then merge alignments following the tree's branching order, until everything is aligned.
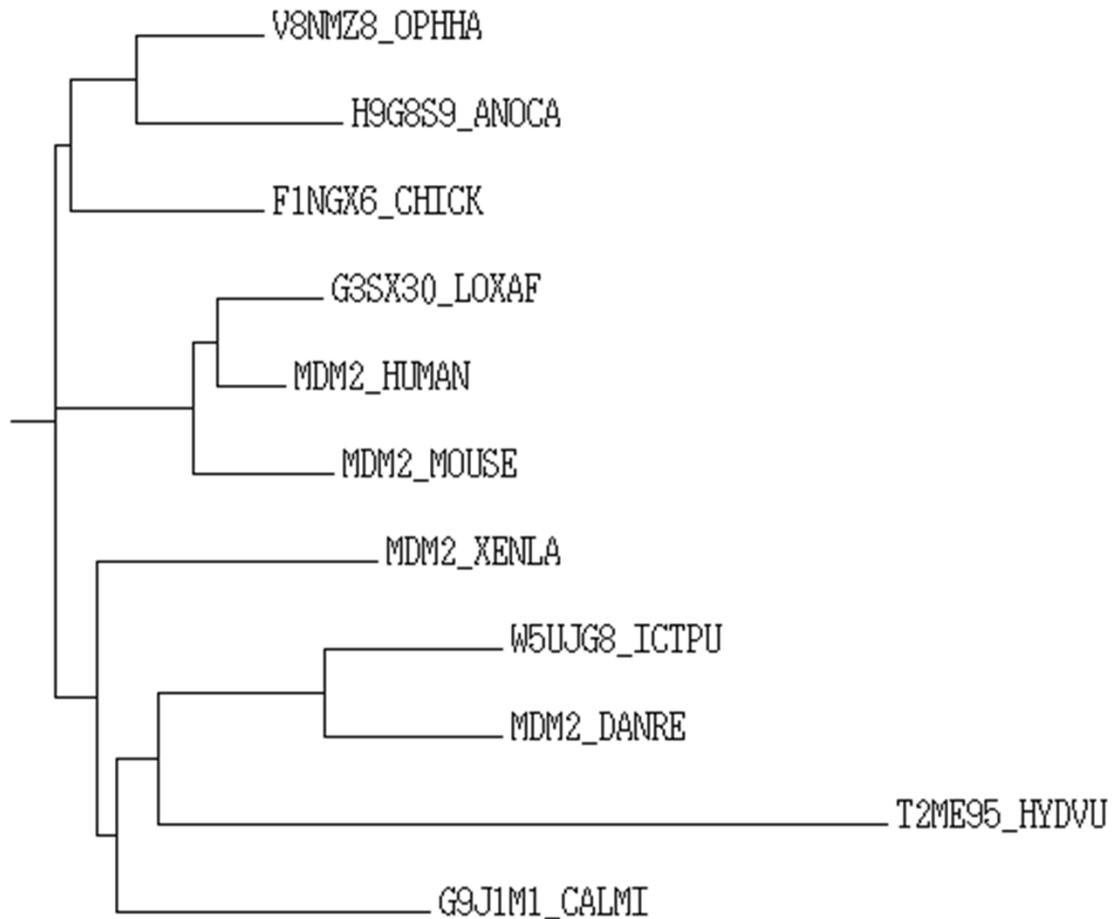
Building Phylogenetic Trees

Practice:

1. Load sequence in FASTA format to Clustal Omega to perform alignment.

2. Download the tree file save in your GMS6014/XXX folder.

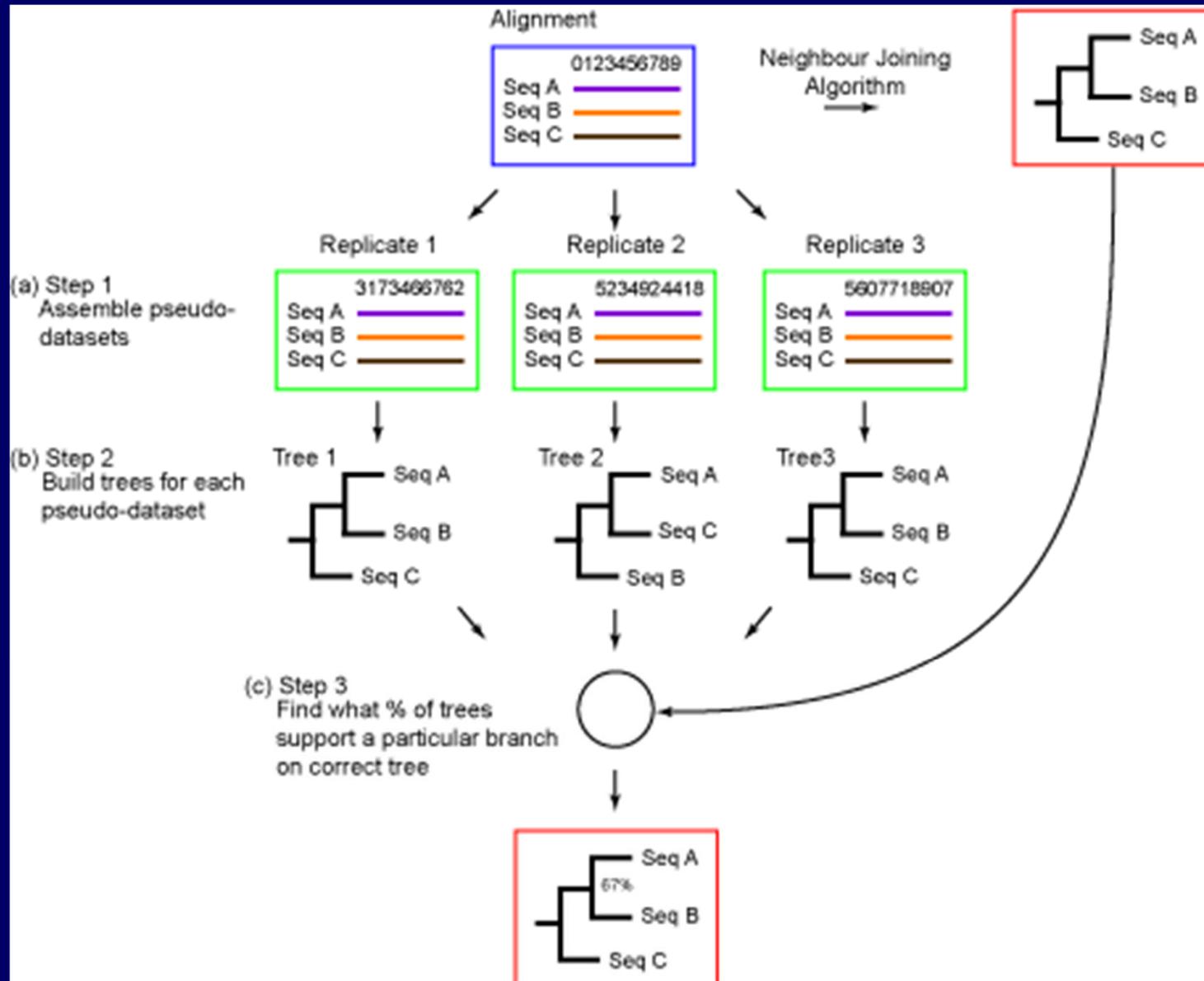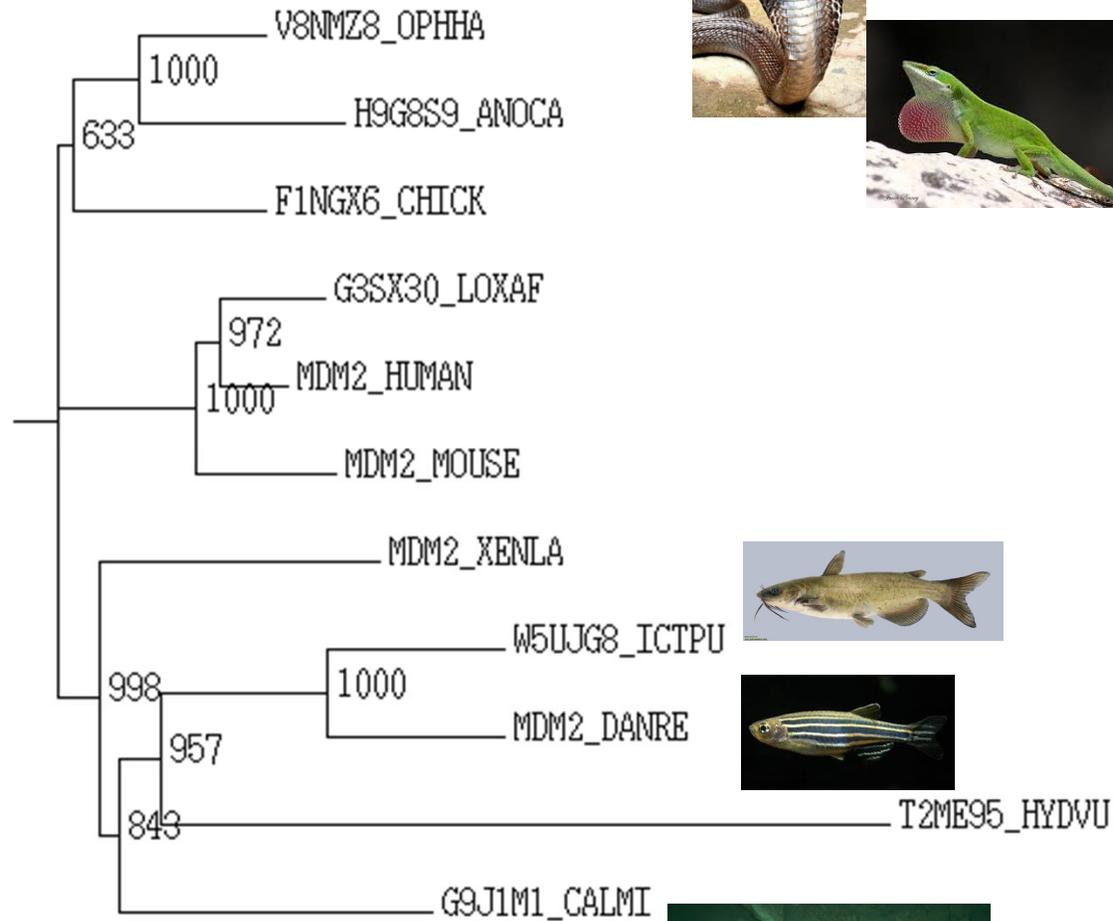3. View the tree with the Phylodendron tree printer.

Phylogenetic Trees

LOXAF: African Elephant

Bootstrapping

Phylogenetic Trees – boot strap

Phylogenetic tree

LOXAF: African Elephant

Building Phylogenetic Trees with JalView

Observe:

Extracting sequence, alignment, and tree building with Jalview.

https://www.jalview.org/