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	

Why a BLAST match is refused by the family ?

Position –specific information about conserved domains is IGNORED in single sequence –initiated search

```
BID_MOUSE     SESQEEIIHN IARHLAQIGDEM DHNIQPTLVR
BAD_MOUSE     APPNLWAAQR YGRELRRMSDEF EGSFKGLPRP
BAK_MOUSE     PLEPNSILGQ VGRQLALIGDDI NRRYDTEFQN
BAXB_HUMAN    PVPQDASTKK LSECLKRIGDEL DSNMELQRMI
BimS          EPEDLRPEIR IAQELRRIGDEF NETYTRRVFA
HRK_HUMAN     LGLRSSAAQL TAARLKALGDEL HQRTMWRRRA
Egl-1         DSEISSIGYE IGSKLAAMCDDF DAQMMSYSAH

BID_MOUSE     SESQEEIIHN IARHLAQIGDEM DHNIQPTLVR
sequence X    SESSSELLHN SAGHAAQLFDSM RLDIGSTAHR
sequence Y    PGLKSSAANI LSQQLKGIGDDL HQRMMSYSAH
```

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 %

Protein motif /domain

- Structural unit
- Functional unit
- Signature of protein family

How are they defined?

Scan protein for functional motifs

Practice: Scan the top hits from BLAST or Motif-based search in Prosite.

- The presence of motif is a strong indication of potential functional and regulatory mechanisms.
- How to interpret short and high frequency motifs?

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

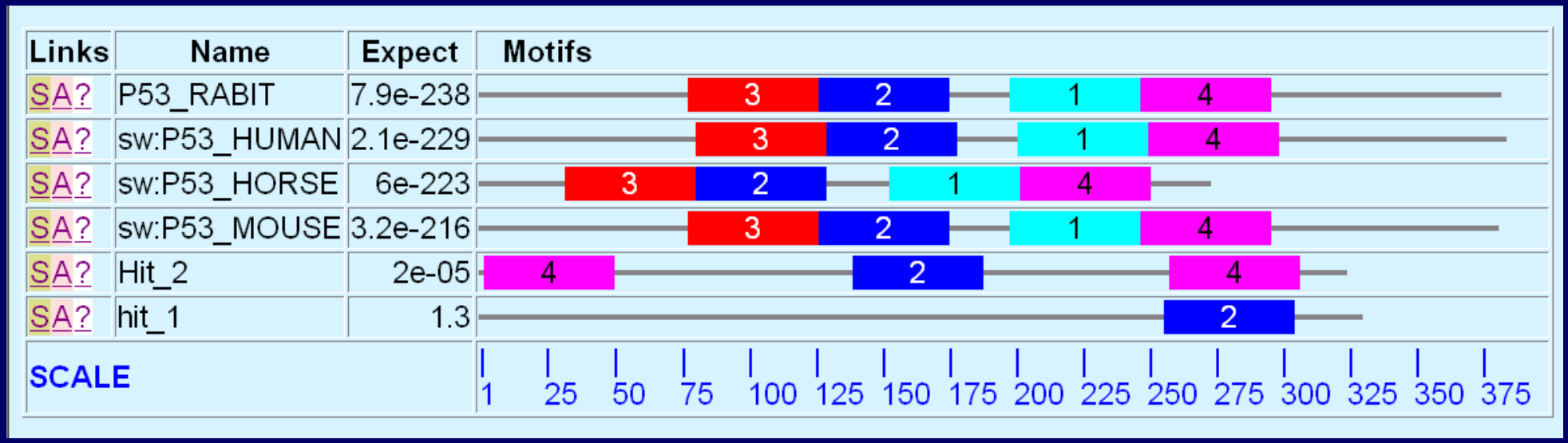
Set1: Mammalian P53 plus mosquito hits

Set2: Diverse set of P53 plus mosquito hits

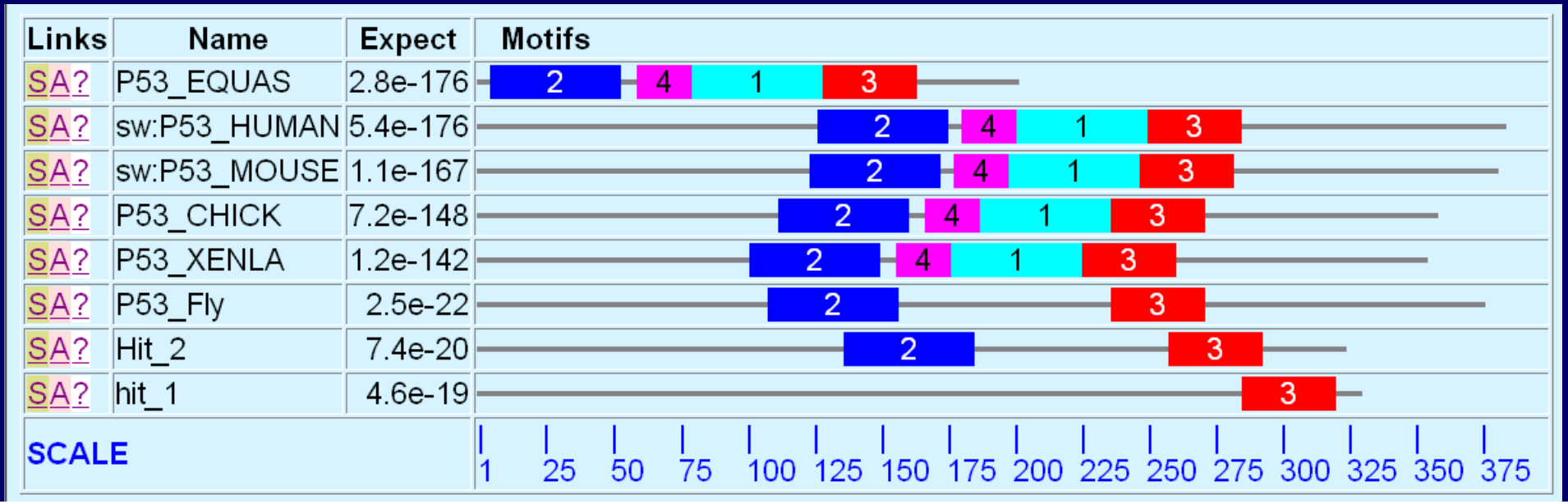
- The outcome of the search is dependent on the inputting set of sequences.
- Compose the inputting set based on your research needs.

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.

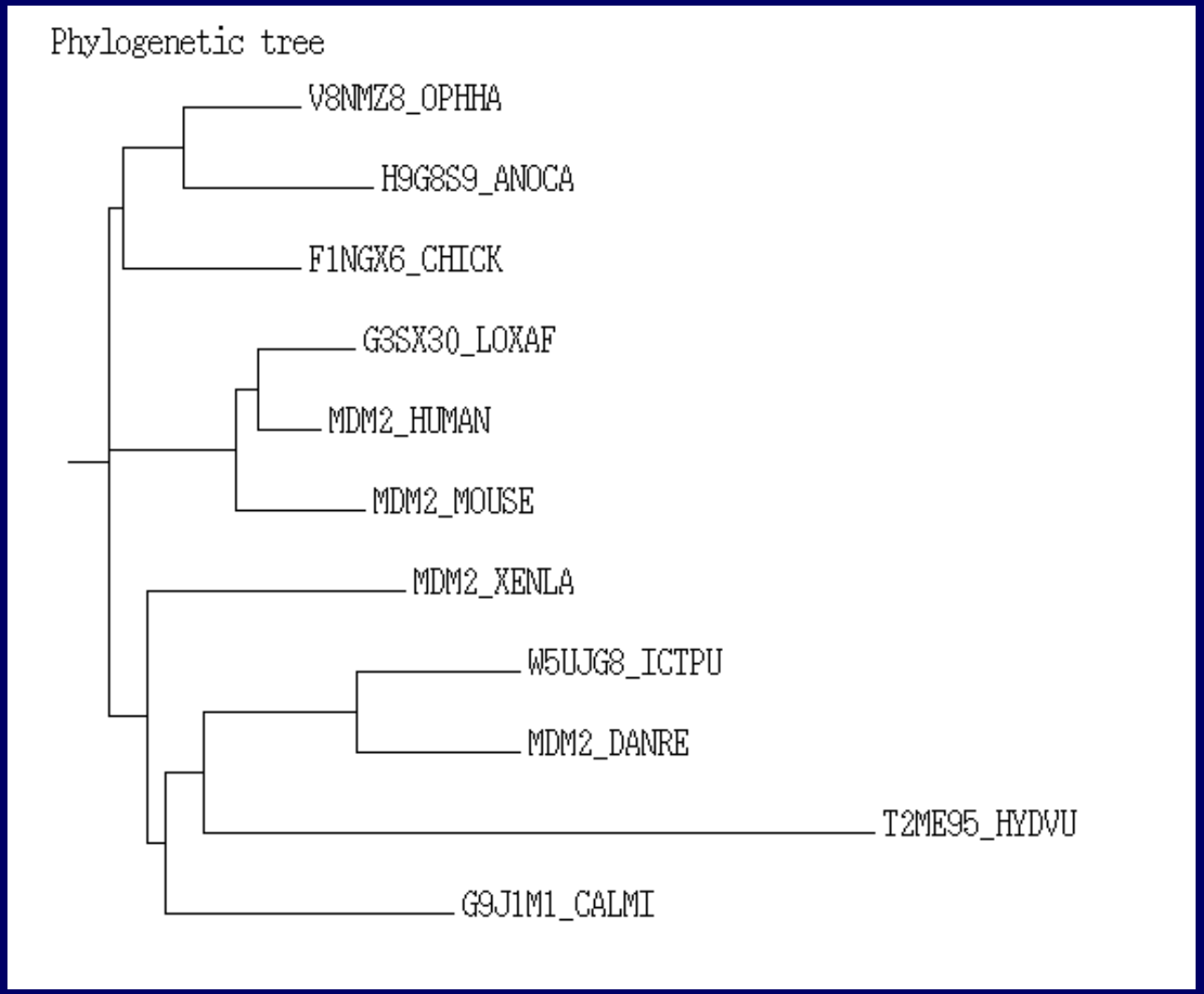
- The distance between sequences.

Building Phylogenetic Trees

Practice:

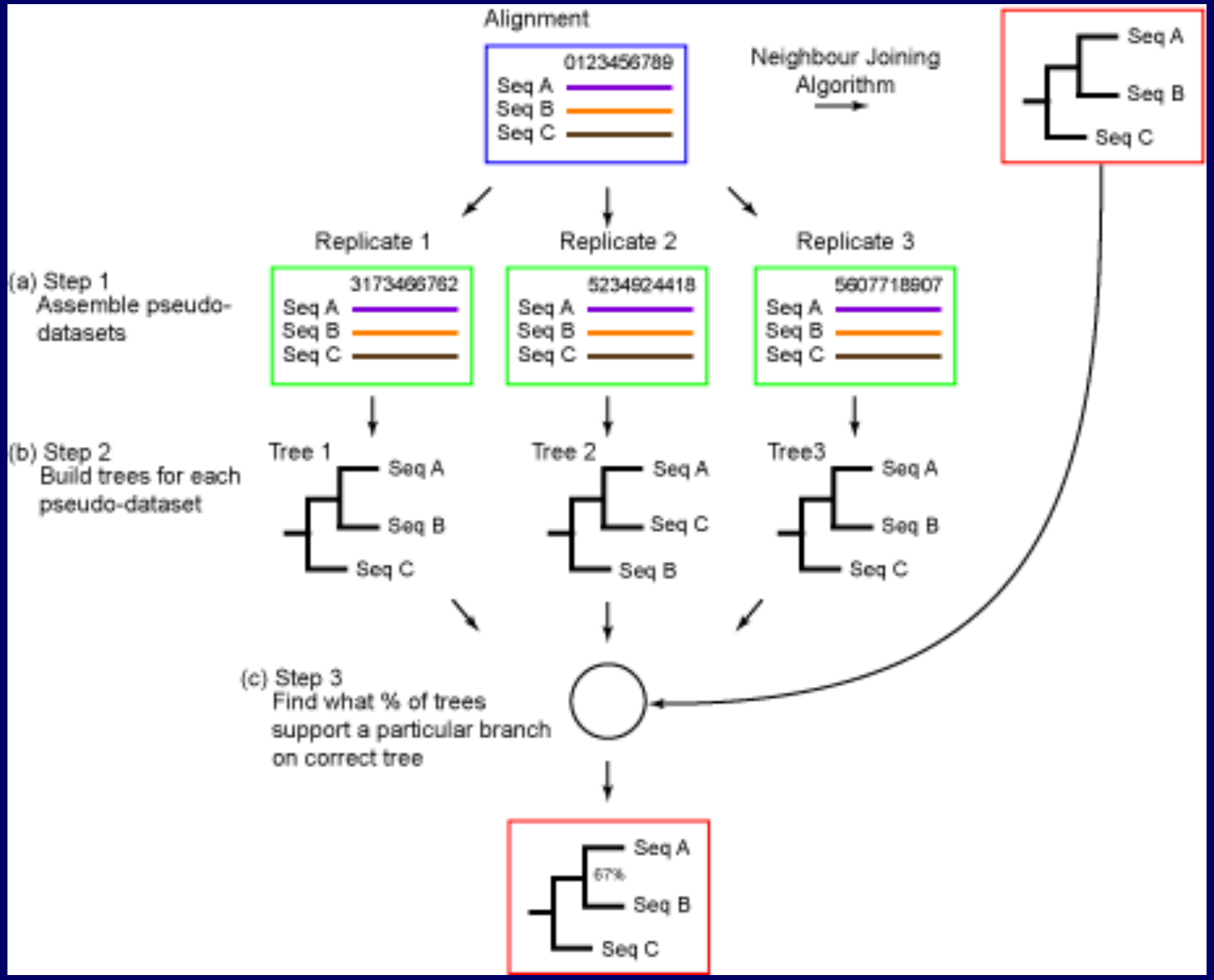
1. Load sequence in FASTA format to ClustalW → perform alignment.
2. Draw Neighbor-joining (N-J) tree.
3. View the tree with the Phylodendron tree printer.
4. Run Boot Strap, view the tree with bootstrap values.

Phylogenetic Trees



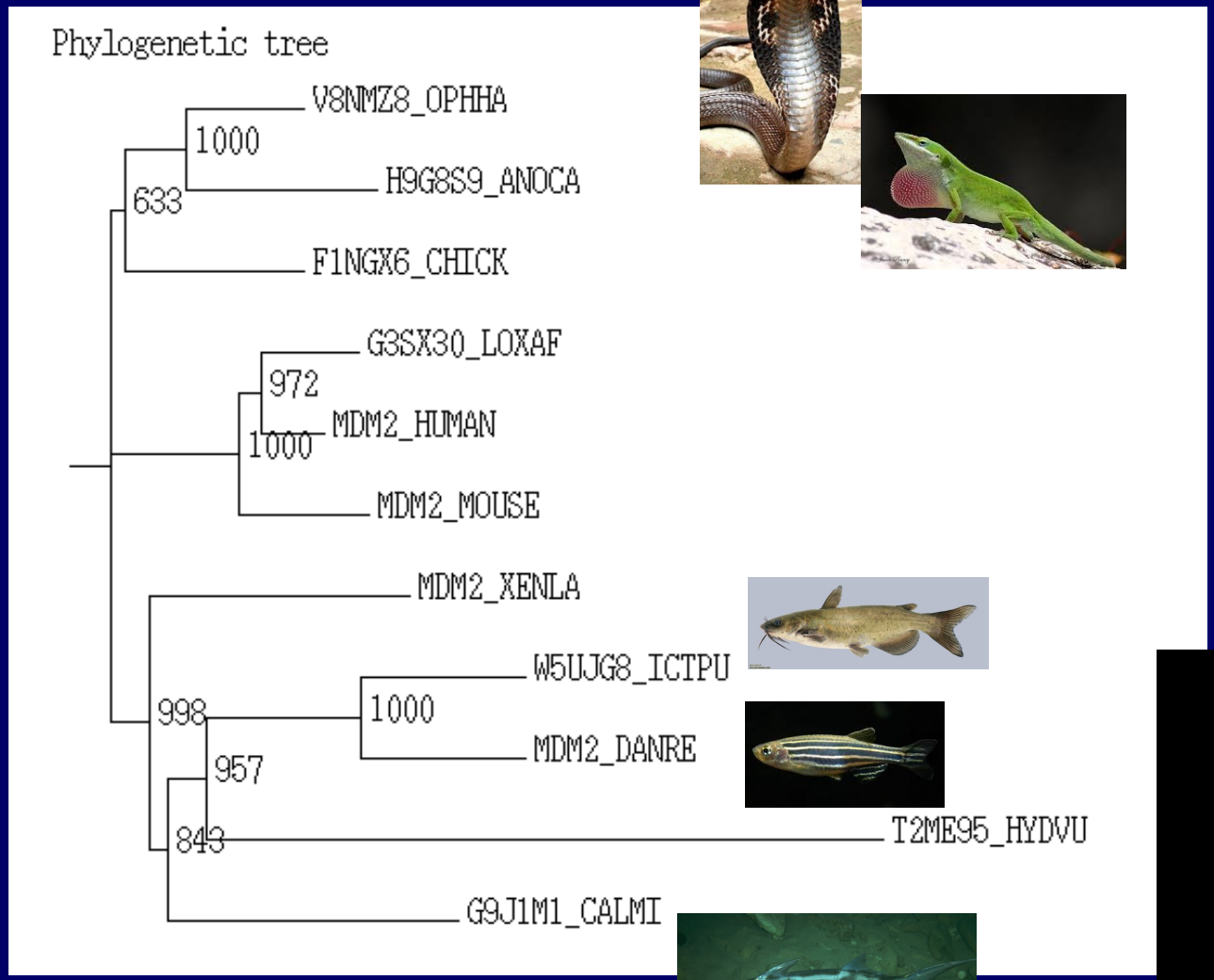
LOXAF: African Elephant

Bootstrapping



http://phylogenetictrees.com/

Phylogenetic Trees – boot strap



LOXAF: African Elephant