## Scoring matrix –BLOSUM 62

			-1 -1 -1 -1 -2		$\begin{array}{c c} 3 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1$	$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c c}                                    $	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	3       1         3       4         3       4         3       -1         3       -1         3       -1         3       -1         3       -1         1       -2         1       -2         1       -2         2       -2         2       -2	1     0     0       1     0     0       1     0       1     1
-2 -1 -1		+1 +1 +2 -2 -1 -1	-1 -1 -1 -1 -1 -1 -1			-1 -1 -1 -1 0	+1 +1 +1 -2 -1 -1 -1 -1 0	-1 -1 -1 -1 -1 -1 -1 -1 -2	10111 -2 -111 -11 -11 -2 -1 -1 -1 -2 -2	+1 +1 +1 +1 +1 +1 +1 -1 +1 -1 +1 -1 +1 -1 +1 -1 +1 -1 +1 -1 +1 -1 +1 -1 +1 -1 +1 -1 -1 -1 +1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
						1 2 2 2 2 1 1 2 1 1 -1 -1 -1 -1 -1 -1			1 + 1 + 2 + 1 + 2 + 1 + 1 + 2 + 1 + 1 +	
	6 1 0 1 0 0	6 1 0 1 0 0	6 1 0 1 0 -2 -3	6 1 0 1 0 -2 -3		6 1 0 1 0 1 0 2 -3 -3 -3 -3			6 1 0 1 0 2 -2 3 -3 -3 -3 -3 -3 -3 -3 -3 -3 -3 -3 -3 -	
6 2 -1 - 2 - 1	6 2 1 - 2 - 1	6 2 -1 -2 -1 -3	6 2 1 - 2 - 3 - 3		6 20 - 1 - 2 - 3 - 3 - 4	62012 	6 2 - 1 - 2 - 3 - 3 - 3 - 3 - 3 - 3 - 3 - 3	520 -1 -2 -3 -4 -3		
5 2 0		5 2 0 1	5 2 0 1 -2 -3	5 2 0 1 -2 -3	5 2 0 1 -2 -3 -3	52 0 1 -3 -3 -2	5 2 0 1 -2 -3 -3 -2	5 2 0 0 1 2 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3		
				1 0 0 H H O 0 0	N O O H H O N	N N 0 0 1 1 1				N H & N N O D H H O O
			8 0 -2 -3		- 2 - 3 - 3	- 2 - 3 - 3 - 3			8 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	
E.	H 2	5 2 -1	5 2 -1 -3	5-1-3	5 -1 -3 -2	ы 1 1 1 3 2 1 3 1 3 1 3 1 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	5 2 -1 -3 -3			
				-1 -3	-1.3	-1				
	5			5	5	51022	51322			
				5	5 1 2	5 1 2	5 1 2 1			
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$
2 5 0 2 5 -1 0 0 8 -2 0 1 0 5 -1 1 -1 2 5	$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	2     0     0     2     5       2     1     -1     0     0     8       2     0     -2     0     1     0     5       2     0     -1     1     -1     2     5	2       0       0       2       5         2       1       -1       0       0       8         2       0       -2       0       1       0       5         2       0       -1       1       -1       2       5         3       -2       -3       -2       0       -2       -1       -1	2       0       0       2       5         2       1       -1       0       0       8         2       0       -2       0       1       0       5         2       0       +1       1       -1       2       5         2       0       +1       1       -1       2       5         3       -2       -3       -2       0       -2       -1       -1         4       -3       -3       -3       -3       -3       -3       -3       -3	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{cccccccccccccccccccccccccccccccccccc$
+1 $+2$ 0 $+2$ 0 1 0 5 -1 $+2$ 0 $-3$ 1 $-1$ 2 5	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$
+1 +2 +1 +2 0 +2 0 1 0 5 -1 -1 -1 -2 6 -1 1 1 -1 2 5	$   \begin{array}{ccccccccccccccccccccccccccccccccccc$	+1 $+2$ $0$ $+2$ $0$ $1$ $0$ $5$ $+1$ $+1$ $+2$ $0$ $+1$ $1$ $-1$ $2$ $5$ $-1$ $+1$ $+2$ $0$ $+1$ $1$ $-1$ $2$ $5$ $-1$ $-2$ $-1$ $-3$ $-2$ $-3$ $-2$ $0$ $-2$ $-1$ $-1$	-1 $-2$ $-1$ $-2$ $0$ $1$ $0$ $5$ $-1$ $-1$ $-1$ $-2$ $0$ $-1$ $1$ $-1$ $2$ $5$ $-1$ $-1$ $-1$ $-2$ $0$ $-1$ $1$ $-1$ $2$ $5$ $-1$ $-2$ $-1$ $-3$ $-2$ $-3$ $-2$ $-1$ $-1$ $-1$ $-3$ $-1$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$	+1 $+2$ $0$ $+2$ $0$ $1$ $0$ $5$ $+1$ $+1$ $+1$ $2$ $0$ $+1$ $1$ $-1$ $2$ $5$ $-1$ $-2$ $-1$ $-1$ $1$ $-1$ $2$ $5$ $-1$ $-2$ $-1$ $-3$ $-2$ $-2$ $0$ $-2$ $-1$ $-1$ $-1$ $-2$ $-1$ $-3$ $-2$ $-3$ $-3$ $-3$ $-3$ $-3$ $-1$ $-3$ $-1$ $-4$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-1$ $-3$ $-1$ $-4$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-1$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-1$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$ $-3$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	-1 $-2$ $-1$ $-2$ $0$ $1$ $0$ $5$ $-1$ $-1$ $-1$ $-2$ $0$ $-1$ $1$ $1$ $-1$ $2$ $5$ $-1$ $-1$ $-1$ $-2$ $-1$ $1$ $1$ $-1$ $2$ $5$ $-1$ $-2$ $-1$ $-3$ $-2$ $-3$ $-2$ $-1$ $-1$ $-1$ $-3$ $-1$ $-4$ $-3$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$
	-1 $-1$ $-1$ $-2$ $0$ $-1$ $1$ $-1$ $2$ $5$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c c c c c c c c c c c c c c c c c c c $	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c c c c c c c c c c c c c c c c c c c $	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$
		1 -1 -2 -1 -3 -2 -3 -2 0 -2 -1 -1	$\begin{array}{cccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{cccccccccccccccccccccccccccccccccccc$

**Basic concept of motif identification 2.** 

#### How do we represent the position specific preference ?

BID_MOUSE	Ι	A	R	H	L	A	Q	I	G	D	E	M
BAD MOUSE	Y	G	R	E	L	R	R	Μ	S	D	E	F
BAK MOUSE	V	G	R	Q	$\mathbf{L}$	A	L	I	G	D	D	Ι
BAXB_HUMAN	L	S	E	C	L	Κ	R	I	G	D	E	L
BimS	I	A	Q	E	$\mathbf{L}$	R	R	I	G	D	E	F
HRK_HUMAN	Т	A	A	R	L	K	A	L	G	D	E	L
Egl-1	Ι	G	S	K	L	A	A	Μ	С	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 <u>model</u>\* Set1: Mammalian P53 plus mosquito hits



#### Set2: Diverse set of P53 plus mosquito hits

Links	Name	Expect	Motifs
SA?	P53_EQUAS	2.8e-176	- <u>2</u> <u>4</u> <u>1</u> <u>3</u> <u>-</u>
SA?	sw:P53_HUMAN	5.4e-176	
SA?	sw:P53_MOUSE	1.1e-167	
SA?	P53_CHICK	7.2e-148	2 4 1 3
SA?	P53_XENLA	1.2e-142	
SA?	P53_Fly	2.5e-22	2 3
SA?	Hit_2	7.4e-20	2 3
SA?	hit_1	4.6e-19	
SCAL	E		I       I

## **Building Phylogenetic Trees**

What is a phylogenetic Tree?

How the observed differences between sequences are developed through evolution.

 $\triangleright$  The distance between sequences.

### **Steps of Building Phylogenetic Trees**

- 1. Collect sequences in one FASTA format file.
- 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.
- 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 -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