Biol5705
Module: Gene Sequence Analysis

Lecture 3
Homology Searching

Dr. Morgan Langille

Outline

- What is a multiple alignment?
- Why do we need a multiple alignment?
- Characters of evolution
- How to create multiple alignments?
- Editing Alignments
- Viewing Alignments

What is a multiple alignment?

- Simply an alignment of more than 2 sequences
- Sequences are aligned globally (end to end)
- Multiple Sequence Alignment (MSA) programs try to insert gaps in the sequences so that *homologous characters* are aligned

What to do with a MSA?

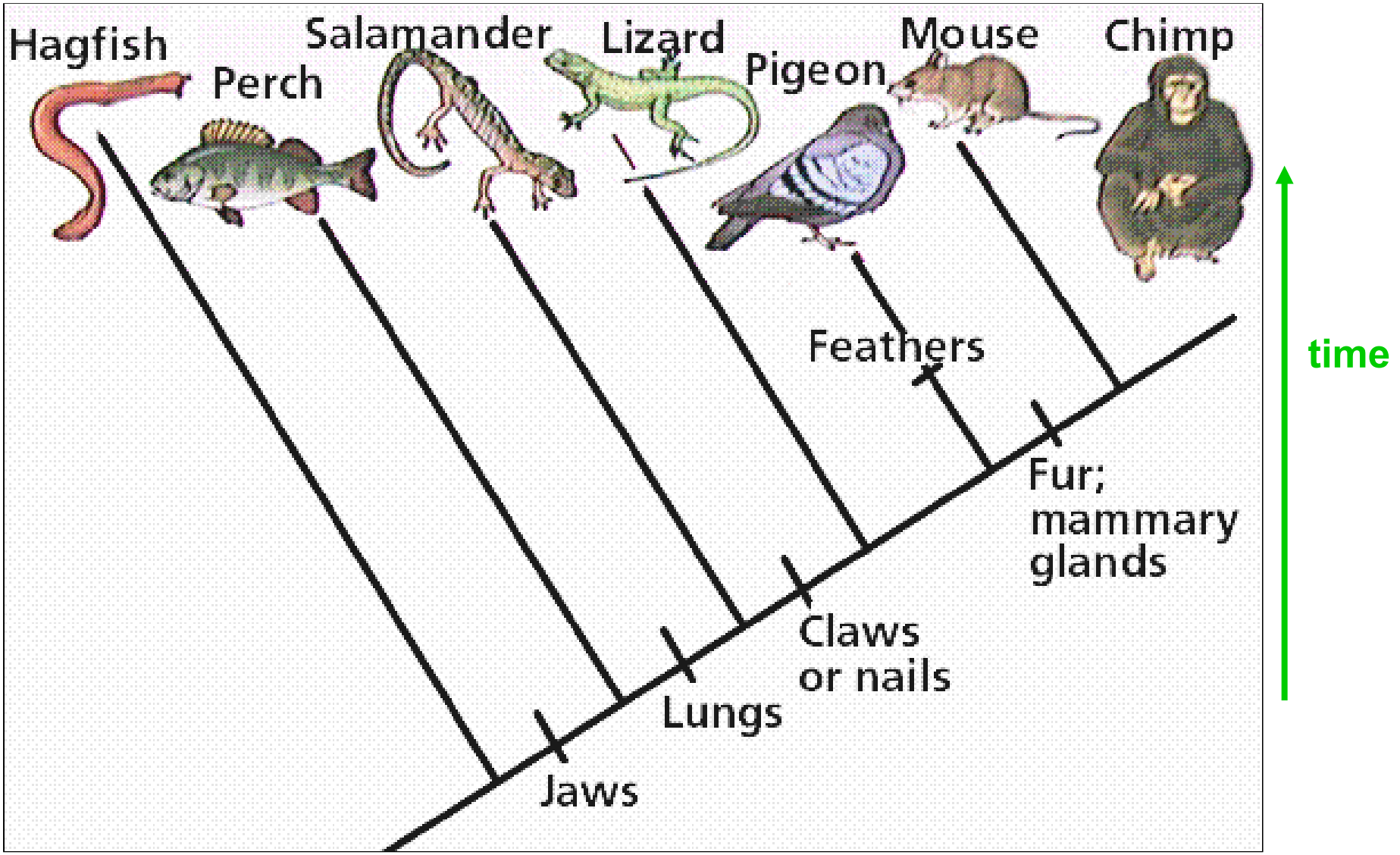
- To build phylogenetic trees
- To look for sites of interest/conservation within a gene (motifs, binding sites, etc.)
- To identify positive selection
  - dN/dS ratios

Characters

- Heritable changes in features (morphology, DNA sequence etc…)

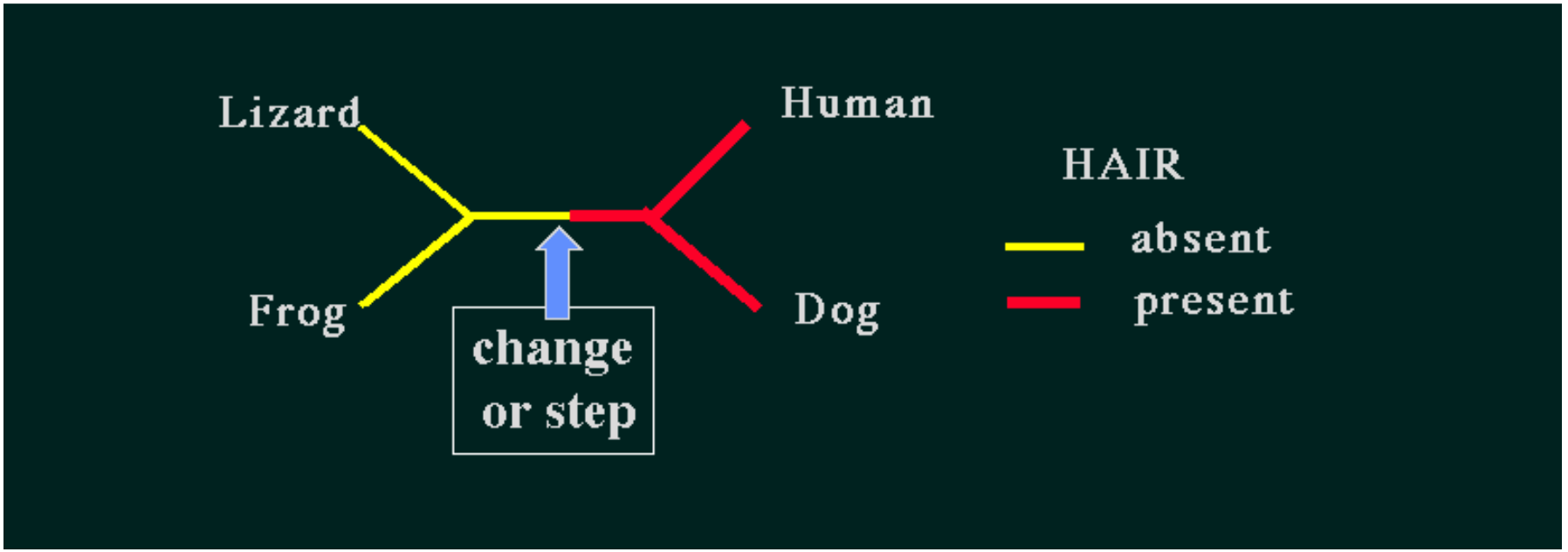
- The more similar characters you have, the more related you are

Evolution and characters



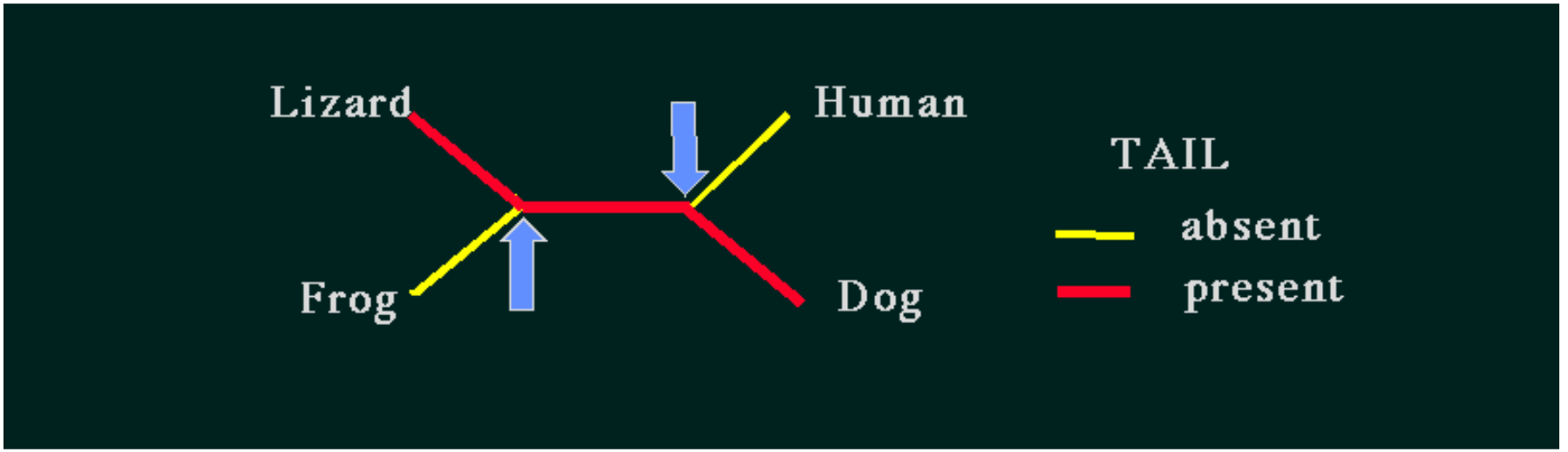
A Unique Character: Hair for Mammals

- Hair evolved only once and is “unreversed”
- Presence of hair → strong indication that organism is a mammal



Homoplasy: The formation of tails

- Tails evolved independently in the ancestors of frogs and humans
- Presence of a tail → no useful conclusions



Classification according to characters – more characters can be good

	Colour	Skin	Cost
Beef	red	no	$$$
Duck	red	yes	$$$
Pork	white	no	$$
Chicken	white	yes	$
Tofu	white	sometimes	$

Chicken most similar to Tofu?

Classification according to characters – increasing the number of characters

	Colour	**Skin**	**Cost**	**Legs**	**Feathers**	**Hair**
Beef	red	no	$$$	four	no	yes
Duck	red	yes	$$$	two	yes	no
Pork	white	no	$$	four	no	yes
Chicken	white	yes	$	two	yes	no
Tofu	white	sometimes	$	none	no	no

Evolution and characters – the importance of comparing characters with common origins (homologous)

bioinformatics

bioinformatics

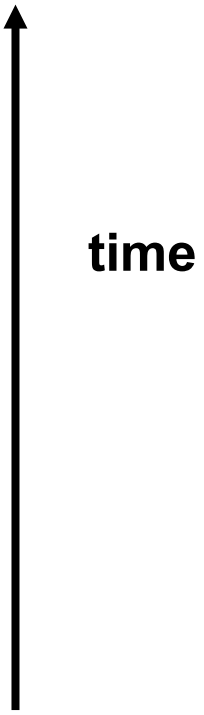
bioinformatios

oinformatios

informatios

information

information



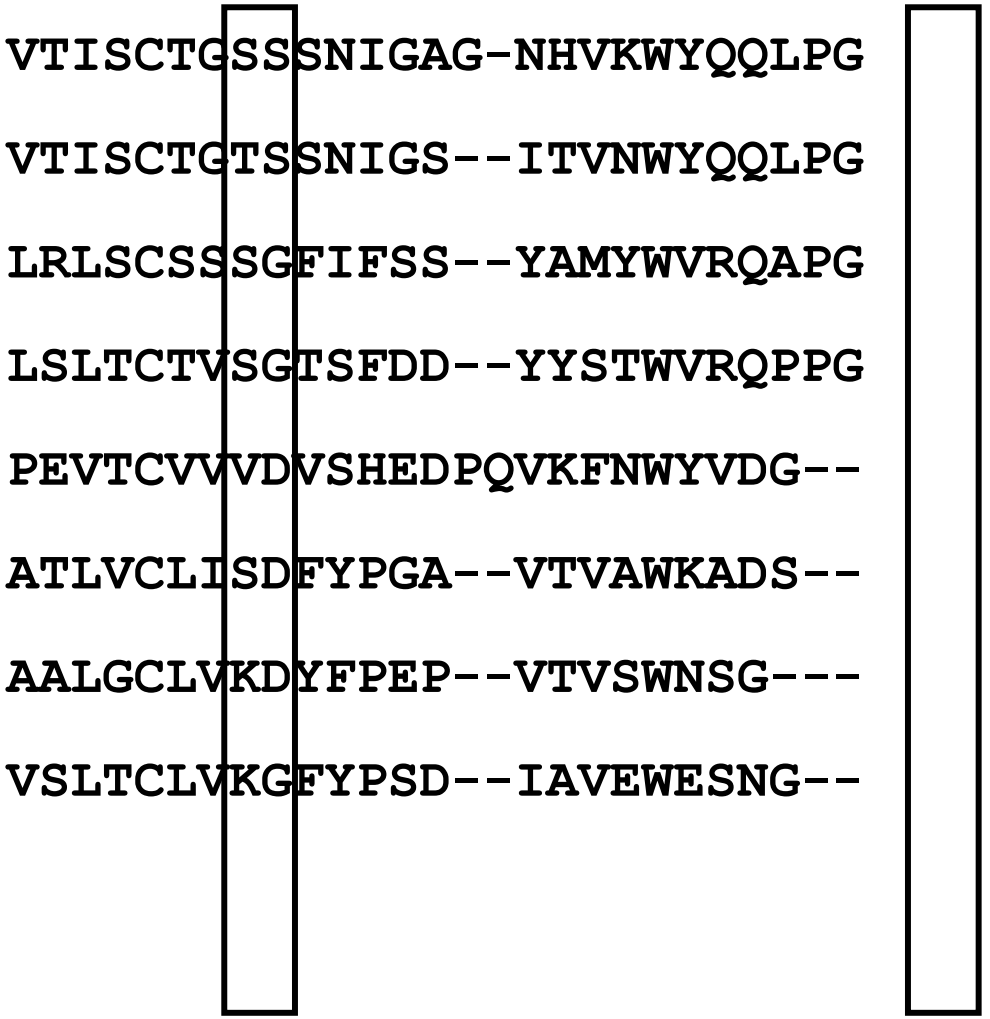
Evolution and characters

```
bioinformatics
bioinformatics
bioinformatios
--oinformatios
---informatios
---information
---information
```

time

- Gaps represent non-homologous positions in the sequence.
- They reflect the occurrence of insertions/deletions or other rearrangements during the evolutionary process.

Multiple Sequence Alignment



The sole purpose of multiple sequence alignments is to place *homologous positions* of *homologous sequences* into the *same column*.

Multiple sequence alignments and phylogenetic analysis

- *First step in any phylogenetic analysis*

- Phylogenetic analysis only as good as the alignment

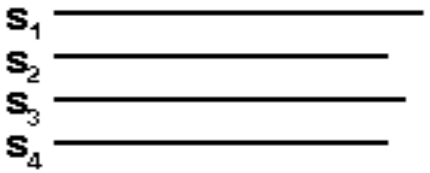
in →

out!

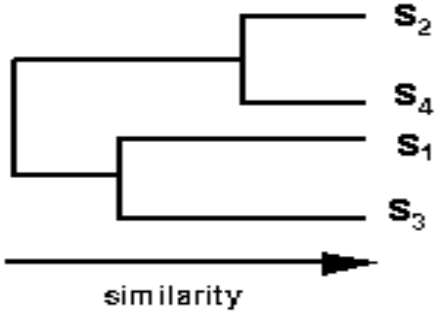
Steps in Multiple Alignment

(A) Pairwise Alignment

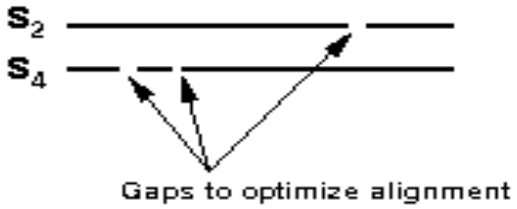
Example – 4 sequences S_1 S_2 S_3 S_4



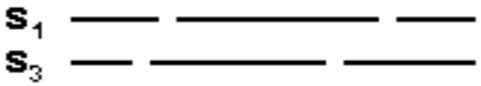
6 pairwise comparisons
then cluster analysis



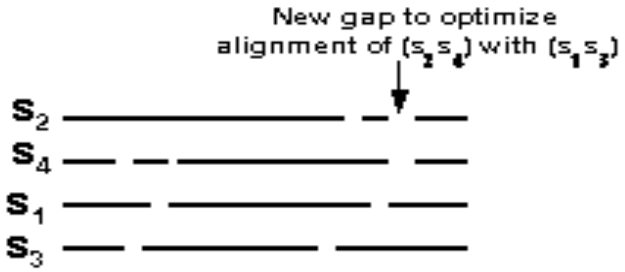
(B) Multiple alignment following the tree from A



align most similar pair



align next most similar pair



align alignments – preserve gaps

Creating a MSA

- Clustal
  - Been around for ever and widely used
  - ClustalW (command line)
  - ClustalX (GUI)
  - Also available on many web servers
  - http://www.clustal.org/clustal2/
- Muscle
  - Faster and maybe more accurate than Clustal
  - Command line only (Although there are web servers)
  - http://www.ebi.ac.uk/Tools/msa/muscle/
- T-Coffee
  - Most accurate, but also the slowest
  - Also has special variations for RNA, protein structure, etc.
  - http://tcoffee.crg.cat/

Need something faster

- Clustal Omega
- HMM Based
- http://www.clustal.org/omega/

Editing Alignments

- A MSA is rarely perfect
- Downstream tools will presume columns are homologous
- Remove unreliably aligned regions for phylogenetic analysis

```
ILPITSPSKEGYESGKAPDEFSSGG
ILPEH--IKDDGELGAAPHSFSTAG
VLPLD-----S--AGRPADSFSAAG
VLPVDR-------DGQARDEYT-VG
VLPVDN-------KGEARDEYT-VG
LLPYDD-------QGRPQDDYSRAG
GIVSRSG---SNFDGEPKDSYGKVG
```

Delete?

Manual vs Automatic

- Manual
  - JalView
- Automatic
  - GBlocks