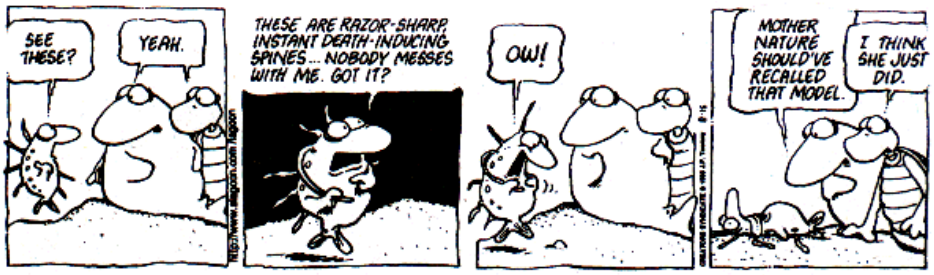
Evolutionary Analysis

Fiona Brinkman
Simon Fraser University,
Greater Vancouver, BC, Canada



Why care about Evolutionary Analysis?

What do

- BLAST
- Protein motif searching
- Protein threading
- Multiple sequence alignment

Have in common?

Why care about Evolutionary Analysis?

Gene family identification

Gene discovery – inferring gene function, gene annotation

Origins of a genetic disease, characterization of polymorphisms

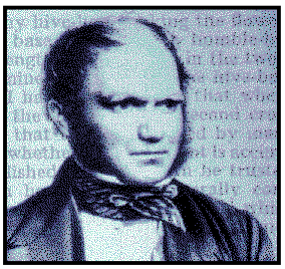
Why care about Evolutionary Analysis?

Koski LB, Golding GB

The closest BLAST hit is often not the nearest neighbor.
J Mol Evol. 2001 Jun;52(6):540-2.

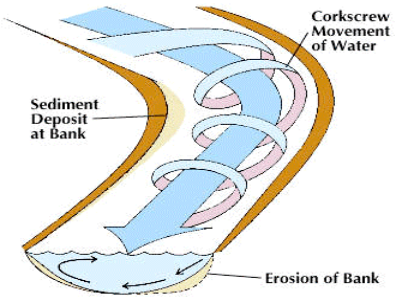
Evolutionary Analysis: Key Concepts

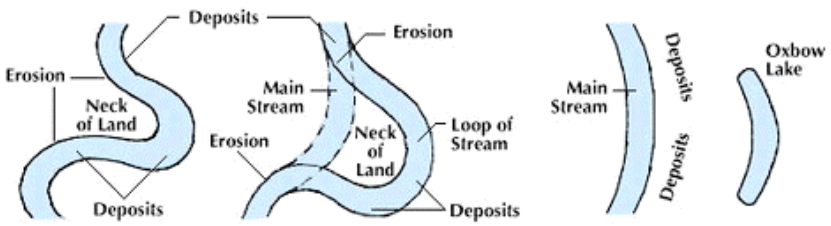
- Foundation of most bioinformatic analyses: Evolutionary theory
- Unique verses non-unique characters
- Sequence alignments are important!
- Fundamentals of phylogenetics and interpreting phylogenetic trees (with cautionary notes)
- Overview of some common phylogenetic methods
- Appreciate the need for new algorithms



18th and 19th centuries: The evolution of a theory

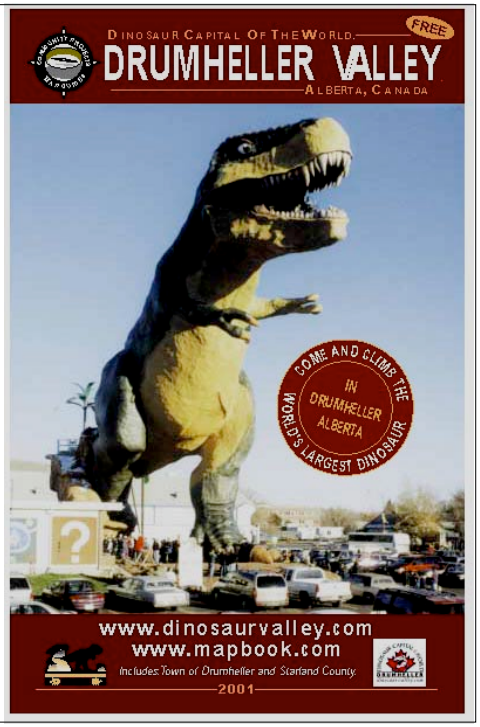
- Earth erosion, sediment deposition, strata – present earth conditions provide keys to the past

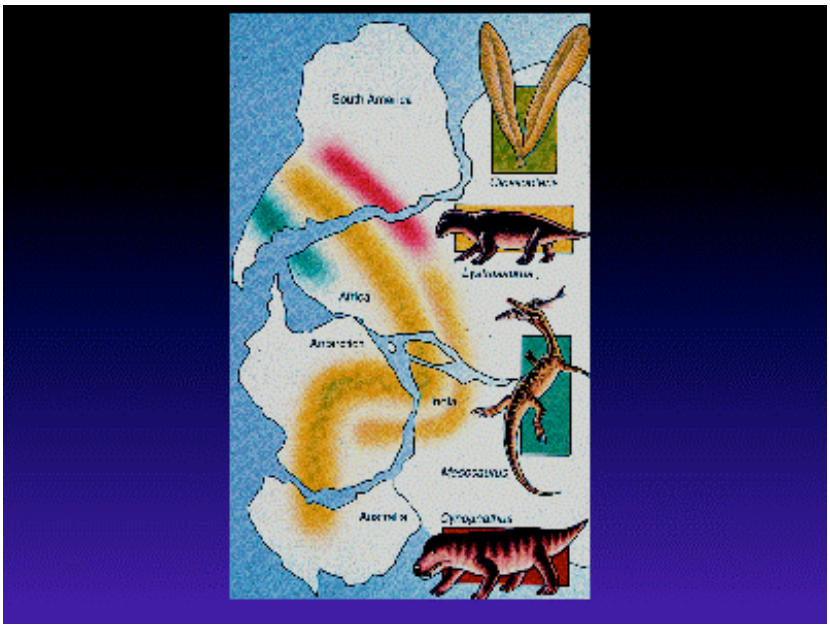




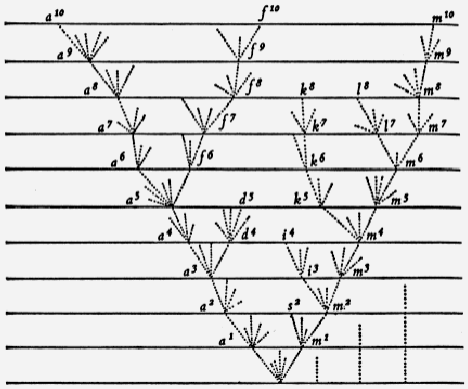
18th and 19th centuries: The evolution of a theory

- Discoveries of fossils accumulated
  - Remains of unknown but still living species that are elsewhere on the planet?
  - Cuvier (circa 1800): the deeper the strata, the less similar fossils were to existing species



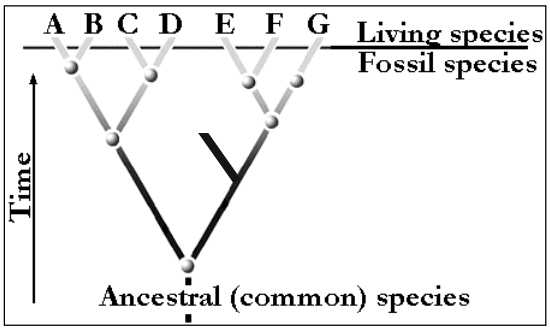


Darwin: "Origin of the species"



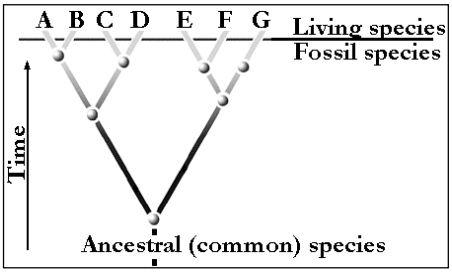
Part of Darwin's Theory

- The world is not constant, but changing
- All organisms are derived from common ancestors by a process of branching.



Part of Darwin's Theory

- This explained…
  - Fossil record
  - Similarities of organisms classified together (shared traits inherited from common ancestor)
  - Similar species in the same geographic region



  - Morphological character-based analysis

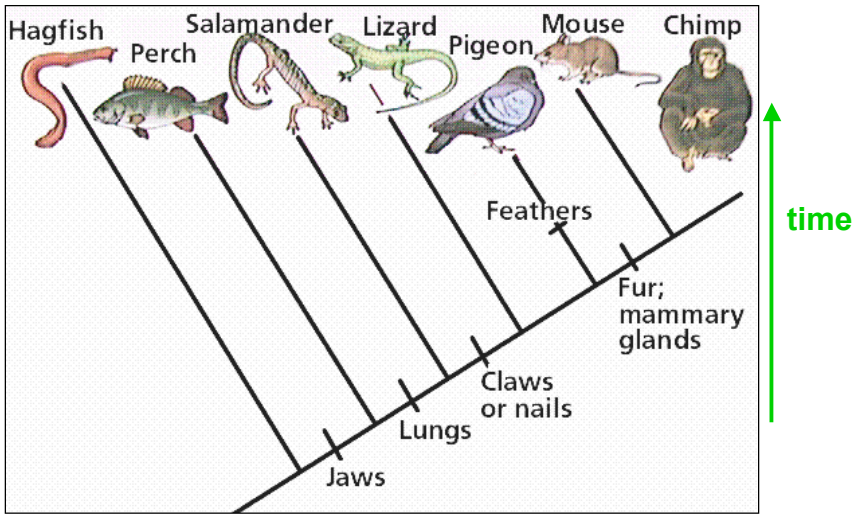
What is evolution?

- Think – Pair – Share!
- Come up with a definition of evolution that is 6 words or less. Bonus points for 2-3 words!

Characters

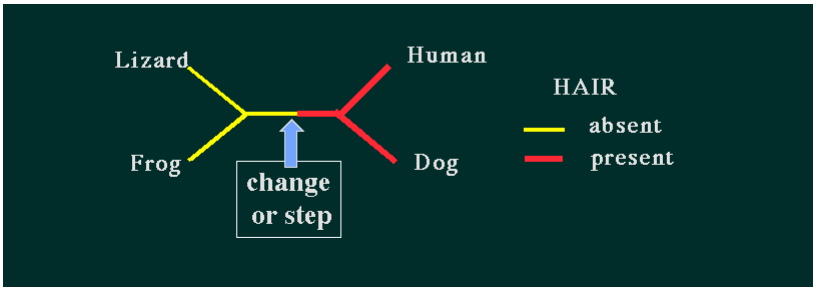
- Heritable changes in features (morphology, DNA sequence etc…)
- The more similar characters you have, the more related you are
- However….. characters can be unique and non-unique

Evolution and characters



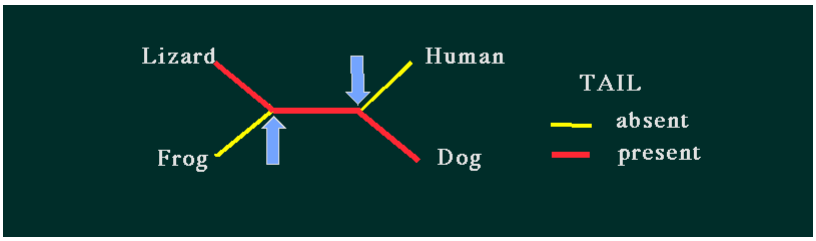
A Unique Character: Hair for Mammals

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

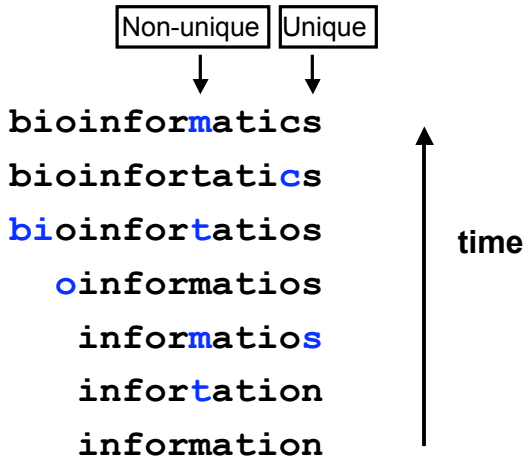


Homoplasy: The formation of tails

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



Unique and non-unique characters



Unique and non-unique characters

Example: Sequence analysis of functionally similar transporters

All share the same deleted sequence region, which is not found in any other transporter examined to date

→Unique character?

→Further investigate for possible functional significance, or use for classification

Unique and non-unique characters

Example: Sequence analysis of functionally similar transporters

All have isoleucine at the third position in the sequence, however some other transporters have isoleucine there too, while some other transporters have leucine at that position

→Non-unique.

→Changes from I → L → I are common (see BLOSUM OR PAM matrices). Not a high priority for further analysis of significance and not useful for classification.

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

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

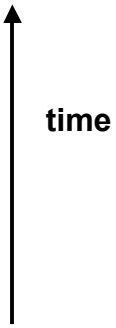
Classification according to characters – increasing the number of characters

	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

Chicken most similar to Duck?

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
```

- 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

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSSSGFIFSS--YAMYWVRQAPG
LSLTCTVSGTSFDD--YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCLISDFYPGA--VTVAWKADS--
AALGCLVKDYFPEP--VTVSWNSG---
VSLTCLVKGFYPSD--IAVEWESNG--
```

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

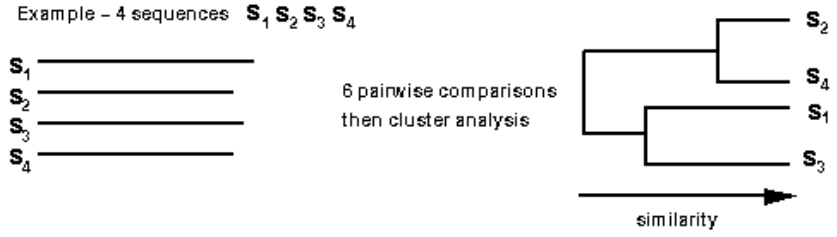
- *First step in any phylogenetic analysis*
- Phylogenetic analysis only as good as the alignment

 in →  out!

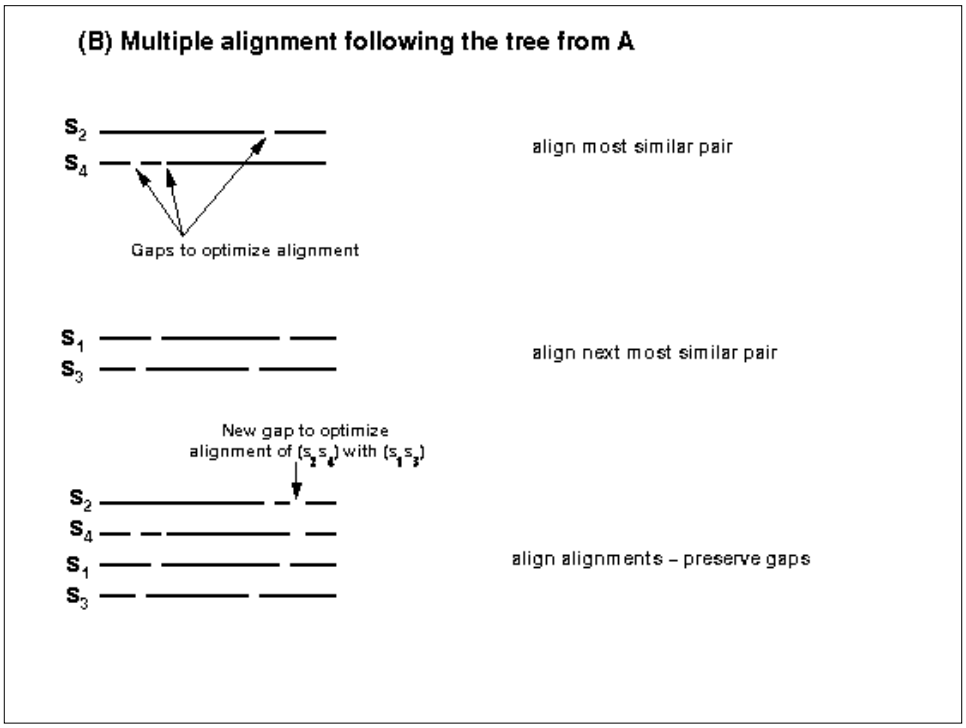
Clustal: Adding evolutionary theory to multiple sequence alignment

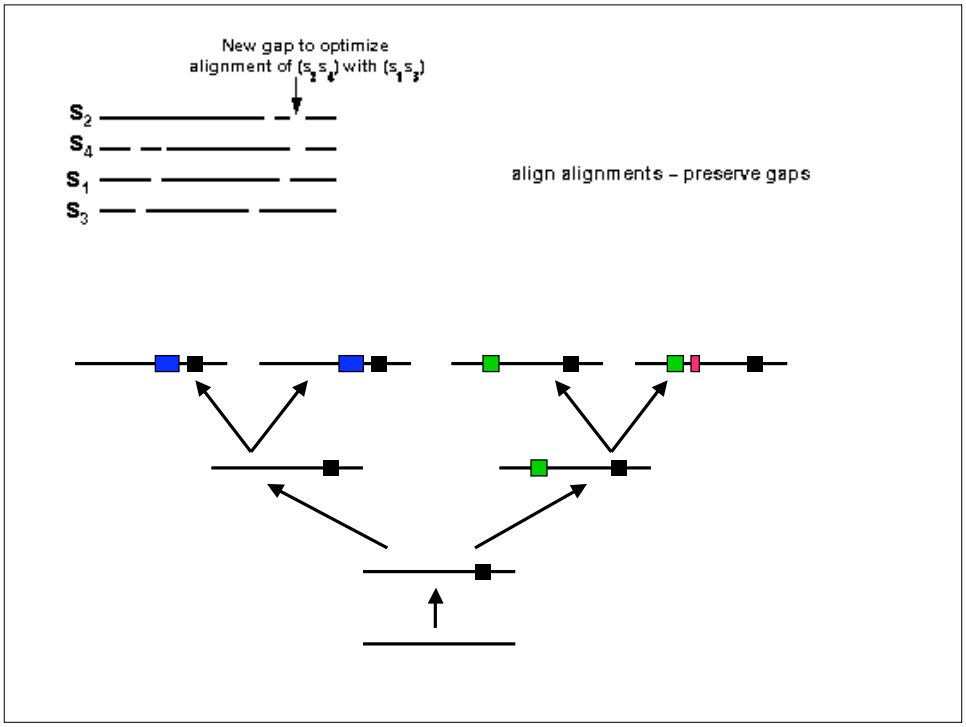
Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22:4673-4680.

(A) Pairwise Alignment



(B) Multiple alignment following the tree from A


(B) Multiple alignment following the tree from A
S2
S4
Gaps to optimize alignment
align most similar pair
S1
S3
align next most similar pair
New gap to optimize
alignment of (s2 s4) with (s1 s3)
S2
S4
S1
S3
align alignments – preserve gaps


New gap to optimize
alignment of (s2 s4) with (s1 s3)
S2
S4
S1
S3
align alignments – preserve gaps

Clustal: Incorporating Biology into Sequence Alignment Algorithms

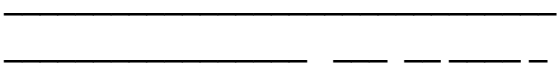
- Matrices varied at different alignment stages according to the divergence of the sequences
- Gap penalties differ for hydrophilic regions to encourage new gaps in potential loop regions
- Gapped positions in early alignments - reduced gap penalties to encourage the opening up of new gaps at these positions

Standard multiple sequence alignment approach (first step for phylogenetic analysis)

- Be as sure as possible that the sequences included are homologous
- Know as much as possible about the gene/protein in question before trying to create an alignment (secondary structure etc..)
- Start with an automated alignment: preferably one that utilizes some evolutionary theory such as Clustal

- Examine alignment:
  - Are you confident that aligned residues/bases evolved from a common ancestor?
  - Are domains of the proteins/predicted secondary structures, etc. aligning correctly?

→ No? May need to edit sequences and redo…



→ Yes? Move on!

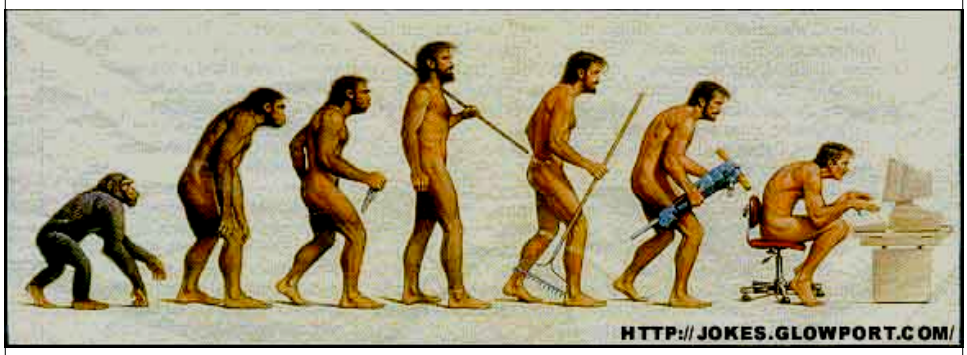
- Note indels (insertions and deletions)
  - Possible insights into functionally important regions…

- Use alignment as a based for subsequent analyses
  (identify consensus or other pattern recognition, for PSSM, HMM construction, phylogenetic analysis, etc..)

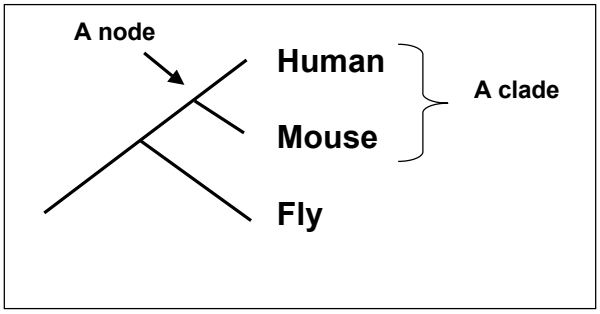
- Remove unreliably aligned regions for phylogenetic analysis

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

Delete?



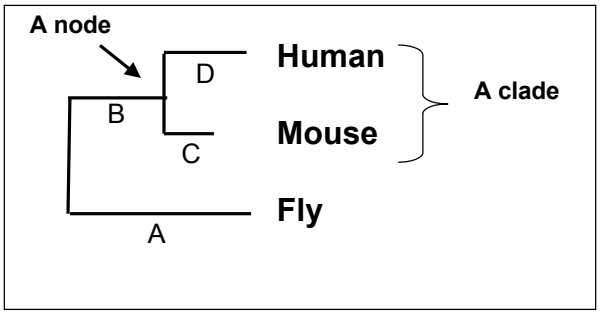
A phylogenetic tree



taxon -- Any named group of organisms – evolutionary theory not necessarily involved.

clade -- A monophyletic **taxon** (evolutionary theory utilized)

A phylogenetic tree with branch lengths



Branch length can be significant…

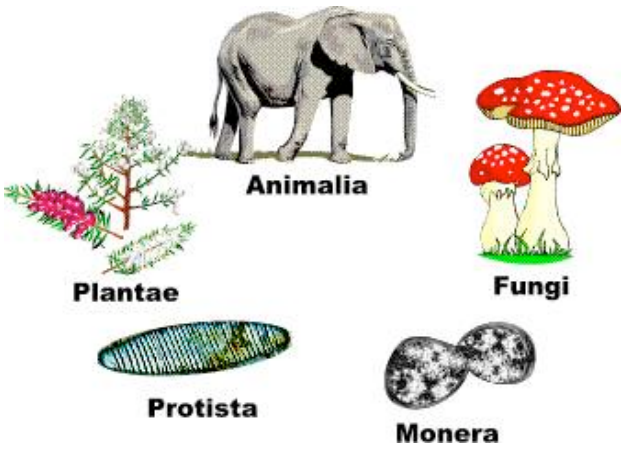
In this case the analysis suggests that the mouse sequence/taxon is slightly more similar to fly than human is to fly

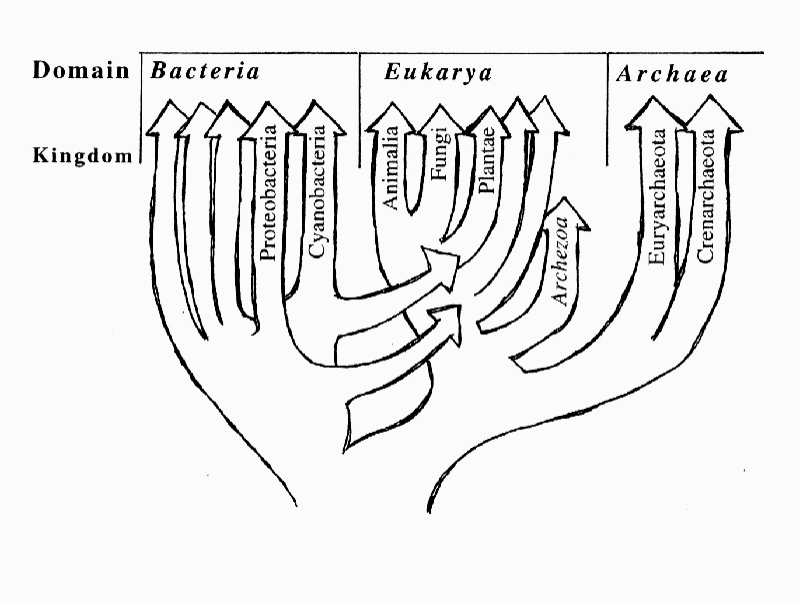
(i.e. sum of branches A+B+C is less than sum of A+B+D)

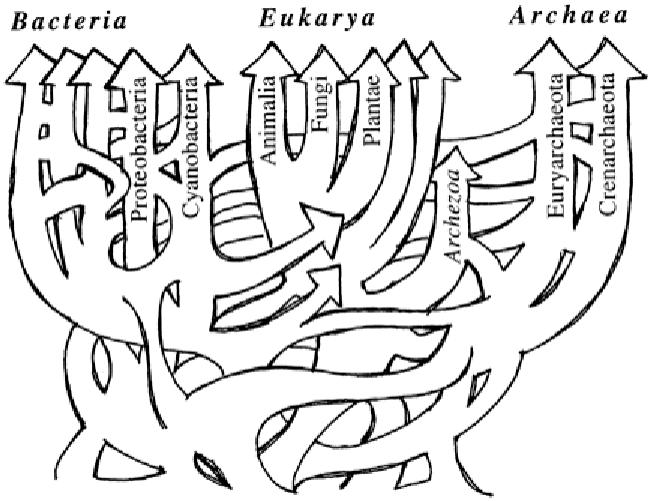
Phylogenetic analysis

- **Organismal relationships**
- **Gene/Protein relationships**

Organismal relationships

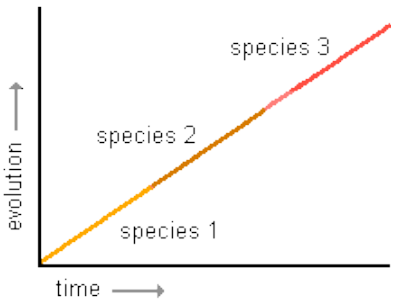


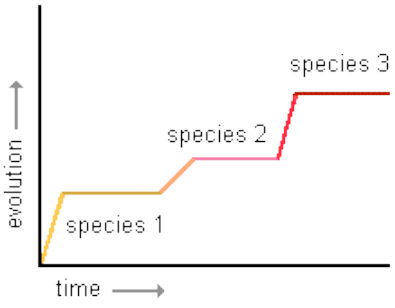




Improving our understanding of organismal relationships

Realization that rates of change are not constant





Improving our understanding of organismal relationships

Better appreciation for what sequences may be suitable for analysis of different degrees of divergence

<u>For the tree of life</u>:

rRNA genes

↓

Multiple genes

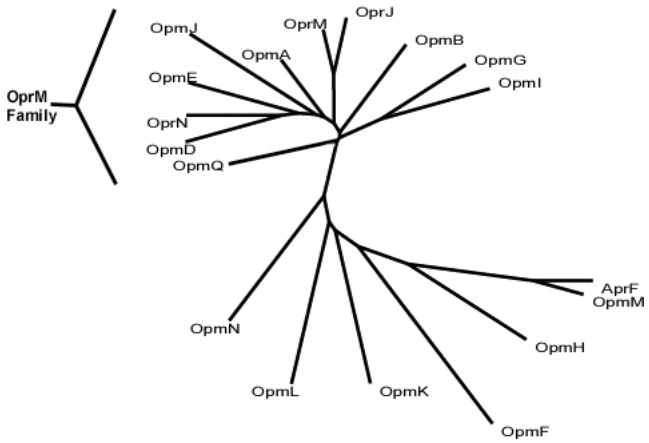
↓

"Whole genome" datasets of genes

↓

rRNA genes and multiple suitable genes

Gene/Protein Relationships



Homolog, ortholog, paralog??

Homologs

Have common origins but may or may not have common activity.

Homologous or not?: *Often* determined by arbitrary threshold level of similarity determined by alignment

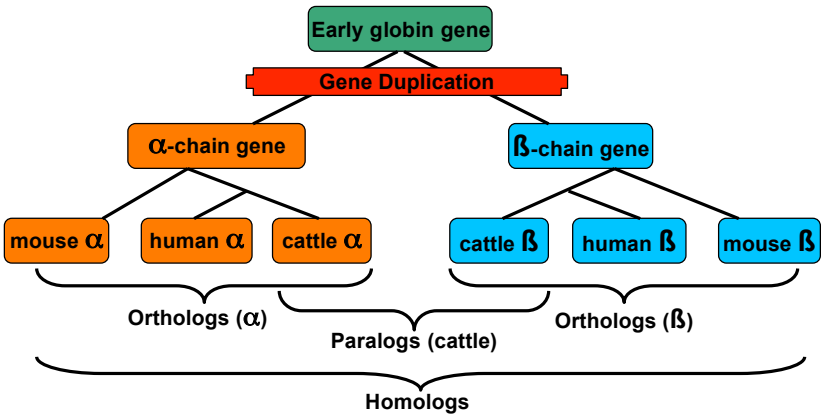
Homologs

…have common ancestry, but the way they are related can vary

(i.e. the reasons they have diverged into different sequences can vary)

- **orthologs** - Homologs produced *only* by speciation. They *tend* to have similar function.
- **paralogs** - Homologs produced by gene duplication. They *tend* to have differing functions.
- **xenologs** -- Homologs resulting from horizontal gene transfer between two organisms.

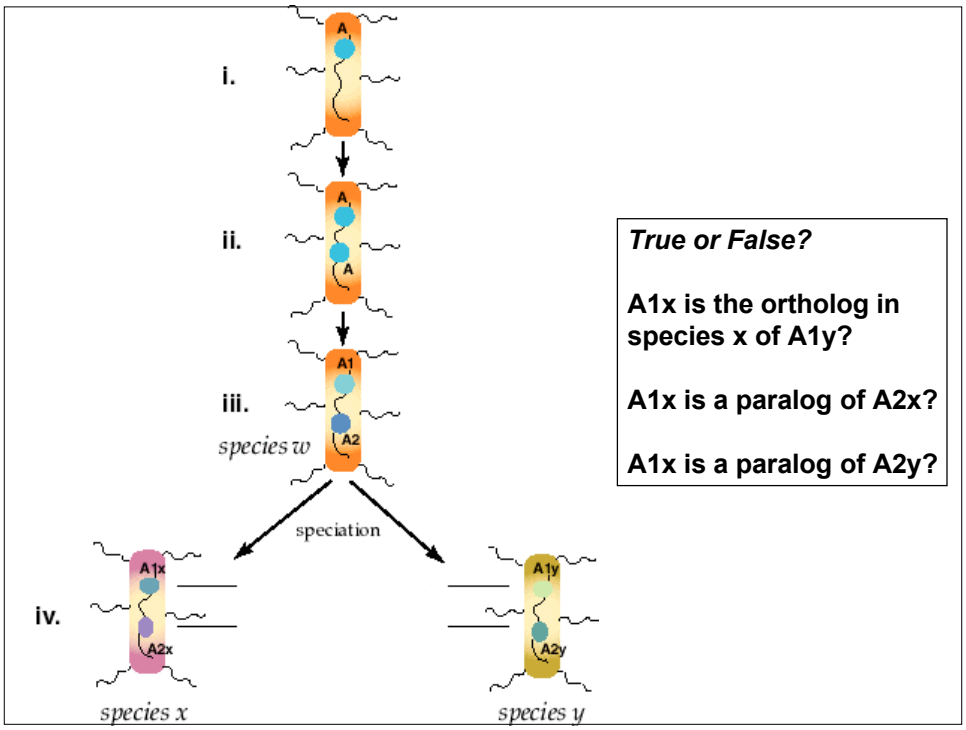
Orthologous or paralogous homologs



Orthologs – diverged only after speciation – *tend to have similar function*

Paralogs – diverged after gene duplication – *some functional divergence occurs*

Therefore, for linking similar genes between species, or performing "annotation transfer", identify orthologs



True or False?

A1x is the ortholog in species x of A1y?

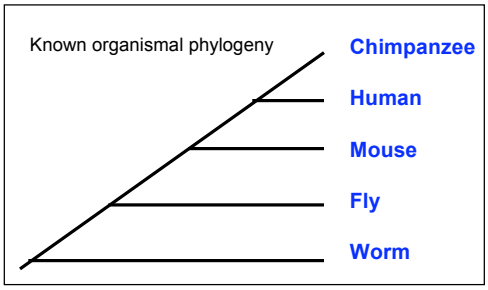
A1x is a paralog of A2x?

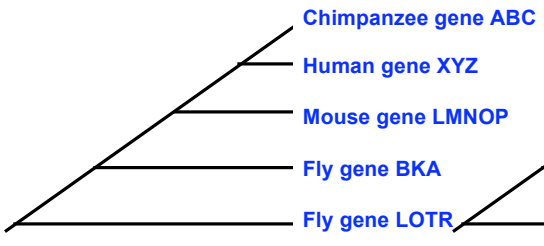
A1x is a paralog of A2y?

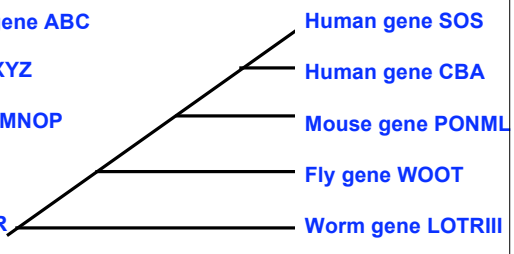
Identifying Gene/Protein Relationships from Phylogenetic trees

- **orthologs** - Homologs produced only by speciation.
  ID: Gene phylogeny matches organismal phylogeny.
- **paralogs** - Homologs produced by gene duplication.
  ID: Multiple copies of homologs in a given species, or genes more/less related than expected by organismal phylogeny.
- **xenologs** -- Homologs resulting from horizontal gene transfer between two organisms.
  ID: Gene phylogeny does not match organismal phylogeny in a tree where most genes do match organismal phylogeny well.

What are the probable orthologs and paralogs of the fly genes BKA and WOOT?





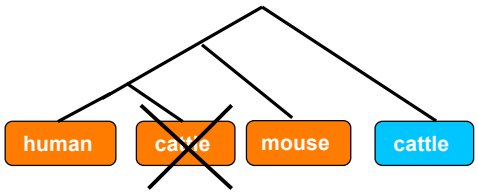


High Throughput Gene Orthology: How to detect?

- Most common high throughput computational method: Identify reciprocal best BLAST hits (EGO, COGs,…)

Example Problem:

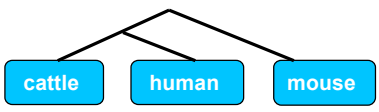
- If making comparisons between human and bovine, for example, the bovine gene dataset is still quite incomplete
- Therefore, current best hit may be a paralog now and the true ortholog not yet sequenced



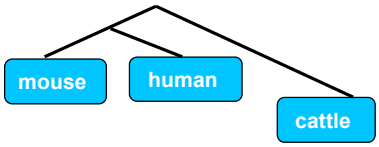
Can we improve orthology analysis for linking functionally similar genes?

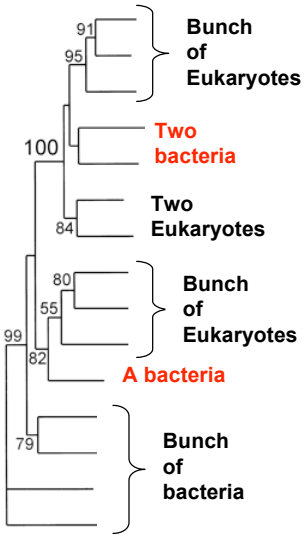
- One solution: Phylogenetic analysis of all putative human-bovine orthologs, using mouse as an outgroup
- Assumption:
  - Mouse and Human gene datasets are more complete, with more true orthologs identified

Expect (organismal phylogeny):

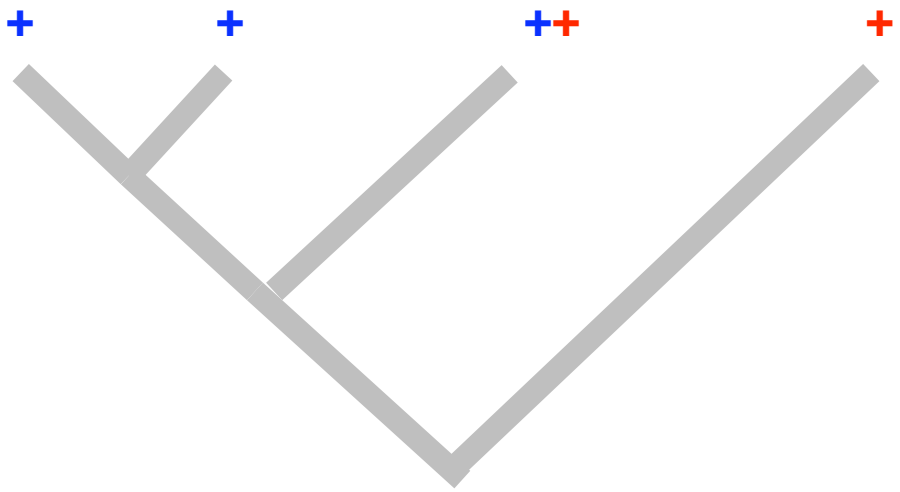


Reject:

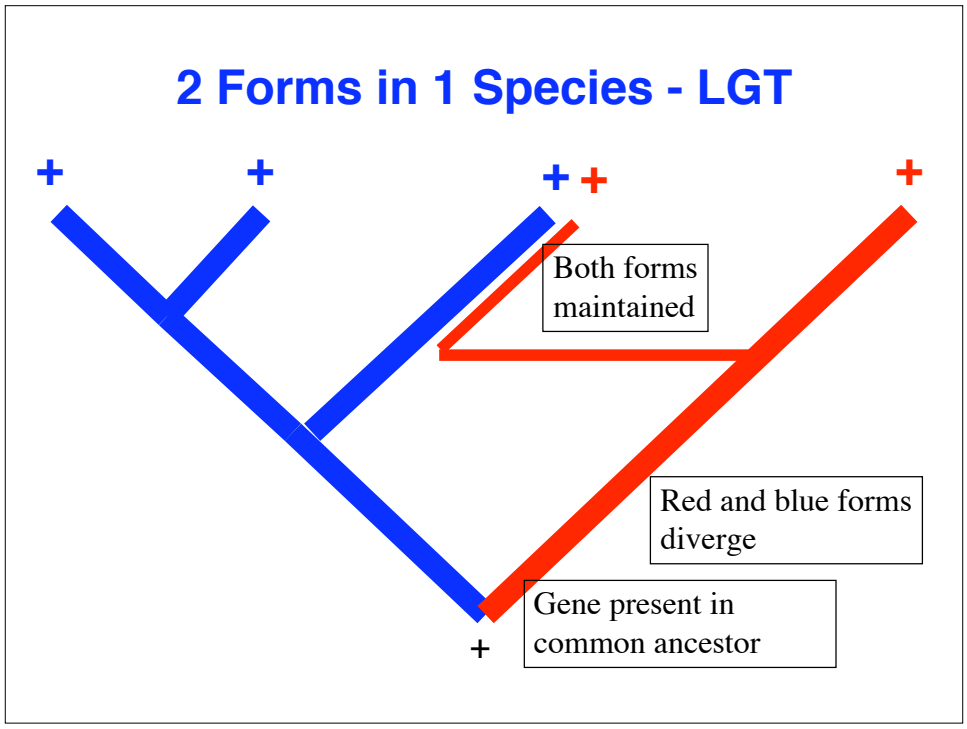


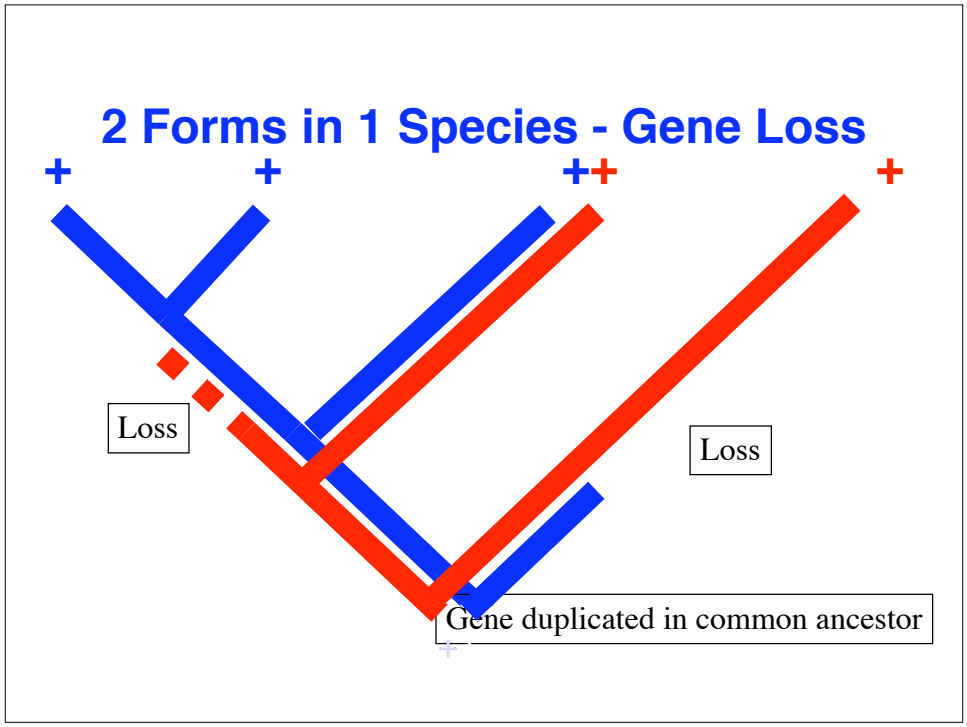


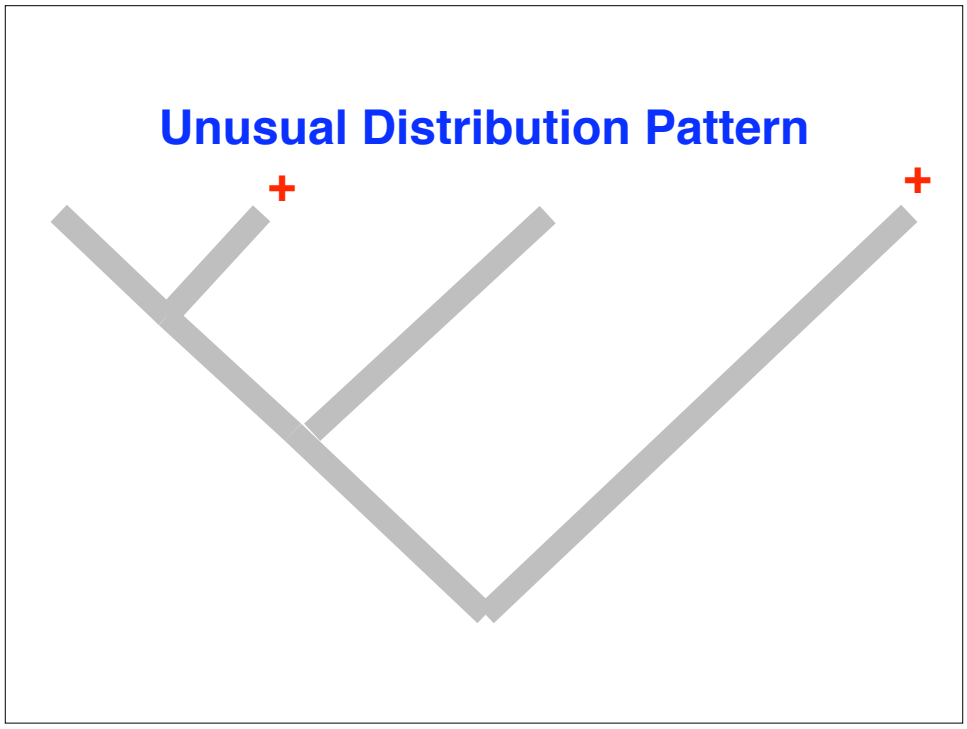
2 Forms in 1 Species

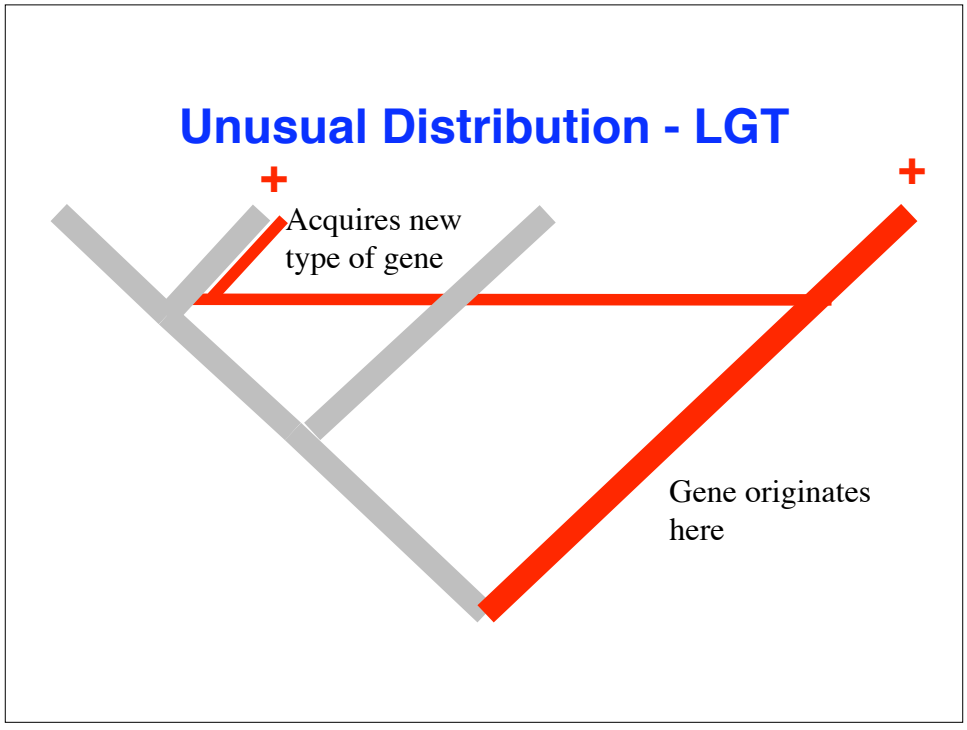


Slides from Jonathan Eisen

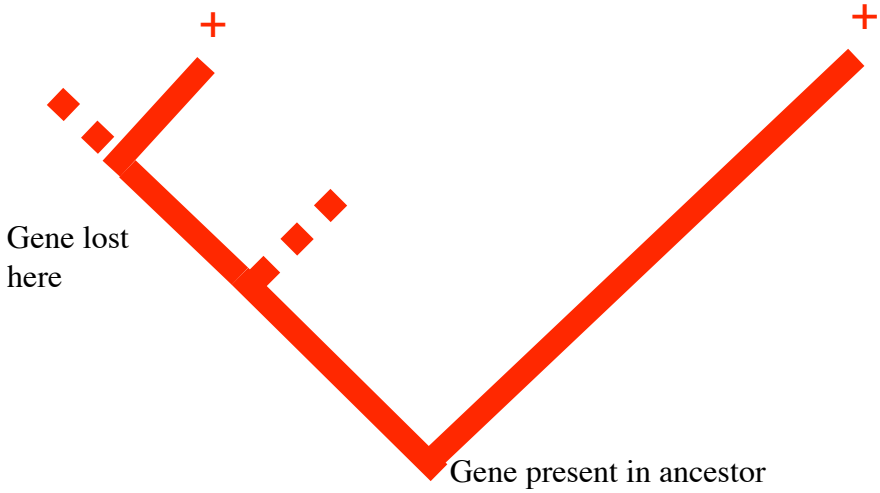

2 Forms in 1 Species - LGT
Both forms maintained
Red and blue forms diverge
Gene present in common ancestor


2 Forms in 1 Species - Gene Loss
Loss
Loss
Gene duplicated in common ancestor

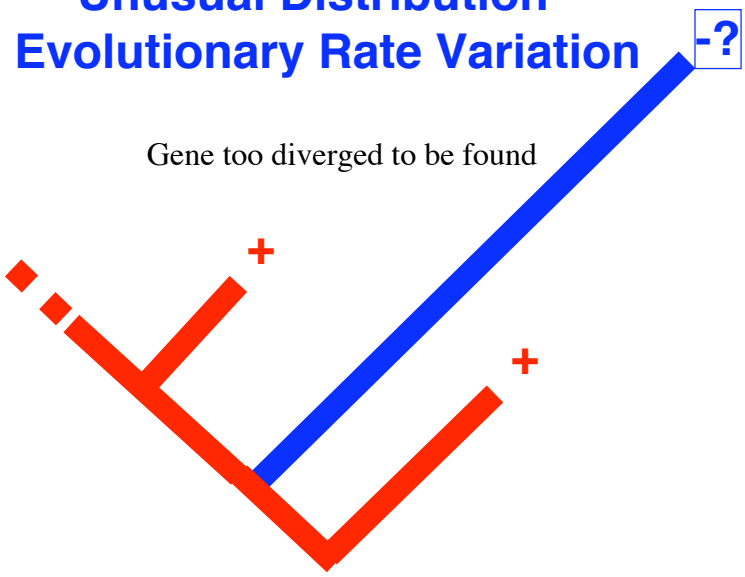

Unusual Distribution Pattern


Unusual Distribution - LGT
Acquires new type of gene
Gene originates here

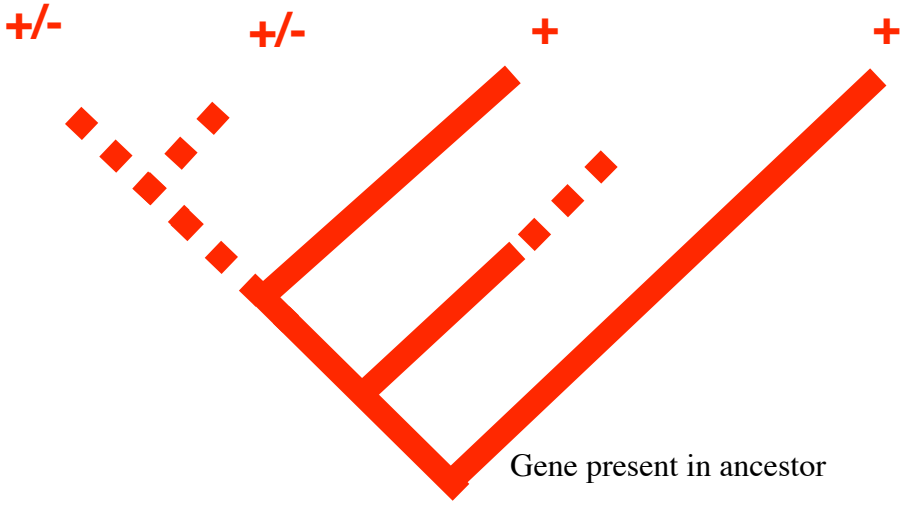
Unusual Distribution - Gene Loss



Unusual Distribution - Evolutionary Rate Variation



Unusual Distribution - Incomplete Data



Hope for the future

Better sampling of all the species in our world

2004: Environmental genomics sampling takes centre stage

Tyson et al (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428, 37-43.

Venter et al (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science,* 304, 66-74.

"So….. how do we construct a phylogenetic tree??"

Most common methods

- Parsimony
- Neighbor-joining
- Maximum Likelihood

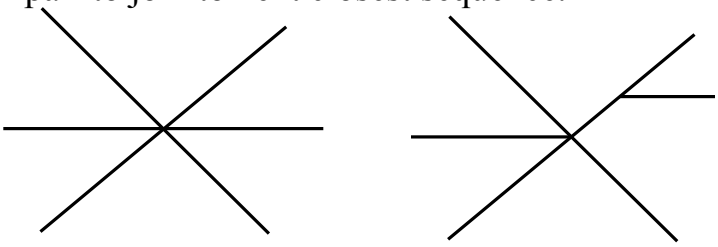
Parsimony

- "Shortest-way-from-A-to-B" method
- The tree implying the least number of changes in character states (most parsimonious) is the best.

- Note:
  - May get more than one tree
  - No branch lengths
  - Uses all character data

Neighbor-joining (and other distance matrix methods)

- "speedy-and-popular" method
- distance matrix constructed
- distance estimates the total branch length between a given two species/genes/proteins
- Neighbor-joining approach: Pairing those sequences that are the most alike and using that pair to join to next closest sequence.



Maximum Likelihood

- "Inside-out" approach
- produces trees and then sees if the data could generate that tree.
- gives an estimation of the likelihood of a particular tree, given a certain model of nucleotide substitution.
- Notes:
  - All sequence info (including gaps) is used
  - Based on a specific model of evolution – gives probability
  - Verrrrrrrrrrry slow (unless topology of tree is known)

How reliable is a result?

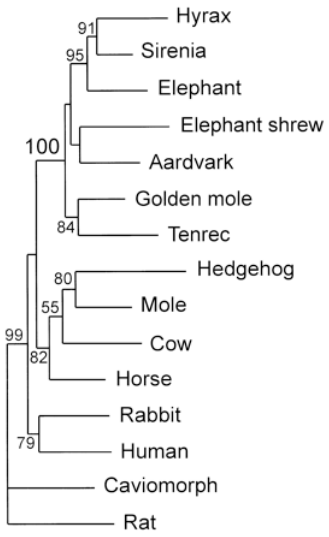
- **Non-parametric bootstrapping**
  - analysis of a sample of (eg. 100 or 1000) randomly perturbed data sets.
  - perturbation: random resampling with replacement, (some characters are represented more than once, some appear once, and some are deleted)
  - perturbed data analysed like real data
  - number of times that each grouping of species/genes/proteins appears in the resulting profile of cladograms is taken as an index of relative support for that grouping

Bootstrapping

The number of times a particular branch is formed in the tree (out of the X times the analysis is done) can be used to estimate its probability, which can be indicated on a consensus tree

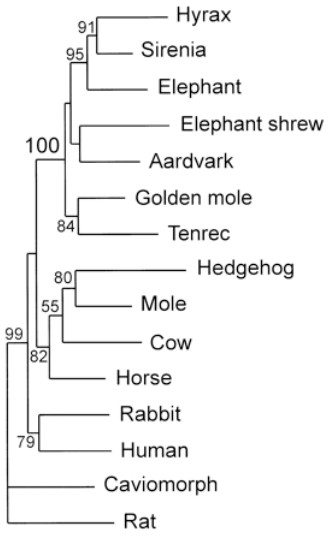
High bootstrap values don't mean that your tree is the true tree!

Alignment and evolutionary assumptions are key



Parametric Bootstrapping

Data are simulated according to the hypothesis being tested.



Phylogenetics – More info

Li, Wen-Hsiung. 1997. Molecular evolution Sunderland, Mass. Sinauer Associates.

- a good starting book, clearly describing the basis of molecular evolution theory. It is a 1997 book, so is starting to get a bit out of date.

Nei, Masatoshi & Kumar, Sudhir. 2000. Molecular evolution and phylogenetics Oxford ; New York. Oxford University Press.

- a more recent book, by two very well respected researchers in the field. A bit more in-depth than the previous book, but very useful.

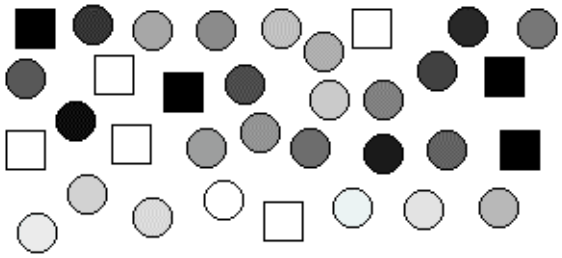
Phylogenetic Tree Construction: Examples of Common Software

PHYLIP
http://evolution.genetics.washington.edu/phylip.html
PAUP
http://paup.csit.fsu.edu/
MEGA 2.1
www.megasoftware.net/

TREEVIEW
http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

Extensive list of software
http://evolution.genetics.washington.edu/phylip/software.html

Challenges



How do we classify?

Computational Challenges

- Need to incorporate more evolutionary theory into the multiple sequence alignment and phylogenetic algorithms used in phylogenetic analysis
- Phylogenetic analyses are computationally intensive – great way to benchmark your CPU speed!
- Automating a continually-updated generation of the Tree of Life, for all genomically sequenced organisms, as more and more genome sequences are determined…

More Challenges

- *Increasing the sampling of our genetic world*
- More accurately differentiating orthologs, paralogs, and horizontally acquired genes
- How frequent is gene loss, gene duplication, and horizontal gene transfer in genome evolution?
- To what degree can we predict protein/gene function using phylogenetic analysis?

Remember:
Evolutionary theory is evolving…

"I've only just bought this bronze stuff and you're telling me I ought to upgrade to iron?"