Evolutionary Analysis

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

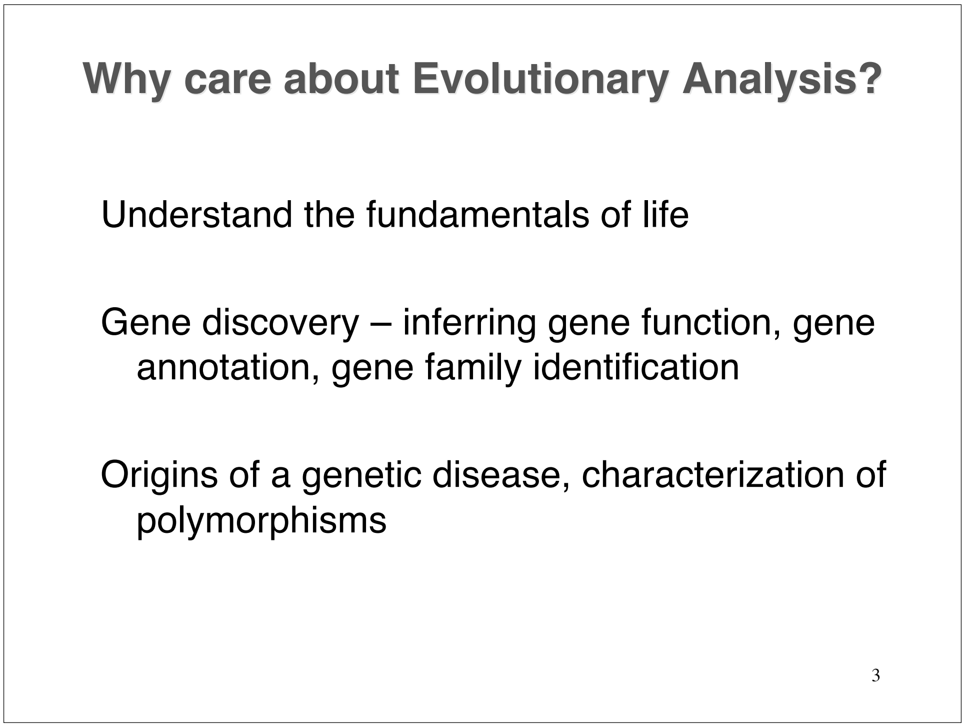
SEE THESE?
YEAH.
THESE ARE RAZOR-SHARP, INSTANT DEATH-INDUCING SPINES... NOBODY MESSES WITH ME. GOT IT?
OW!
MOTHER NATURE SHOULD'VE RECALLED THAT MODEL.
I THINK SHE JUST DID.

Why care about Evolutionary Analysis?

What do

- BLAST
- Protein motif searching
- Multiple sequence alignment

Have in common?

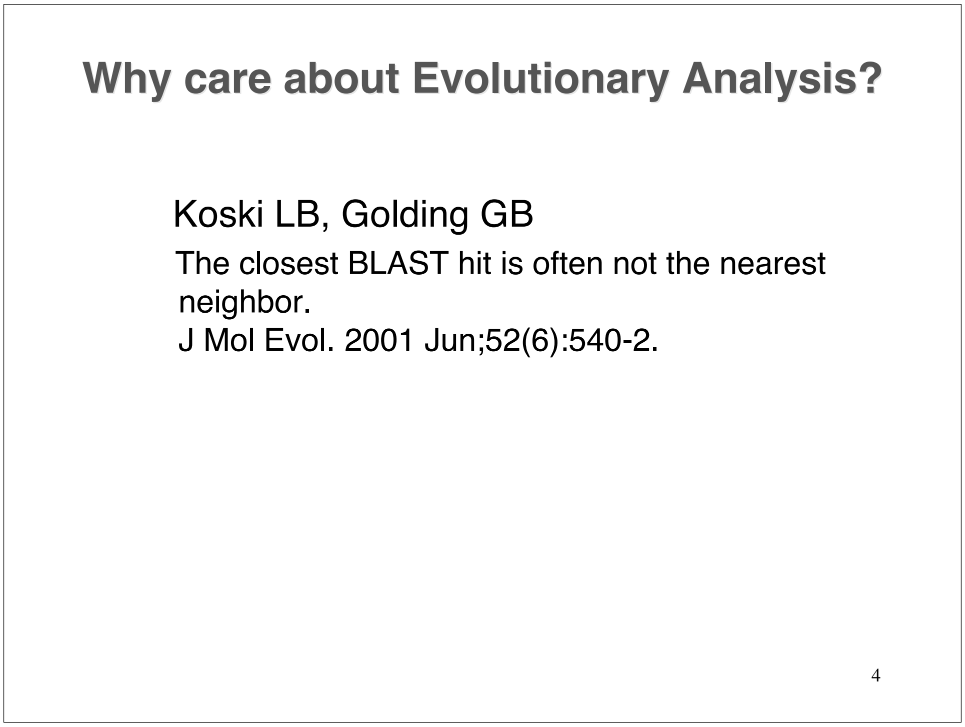


Why care about Evolutionary Analysis?

Understand the fundamentals of life

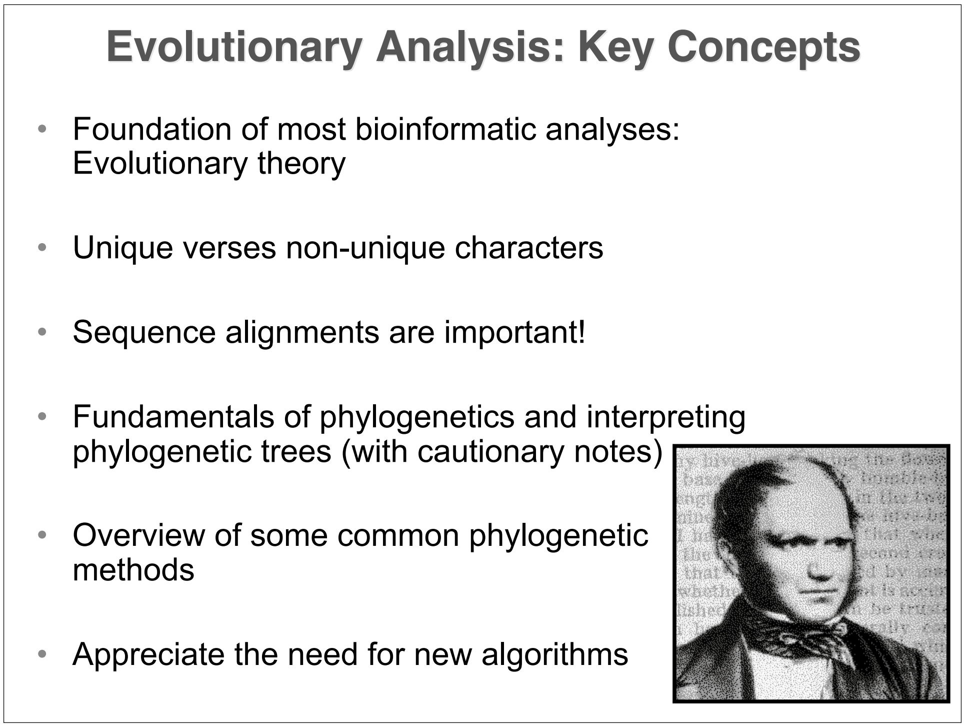
Gene discovery – inferring gene function, gene annotation, gene family identification

Origins of a genetic disease, characterization of polymorphisms



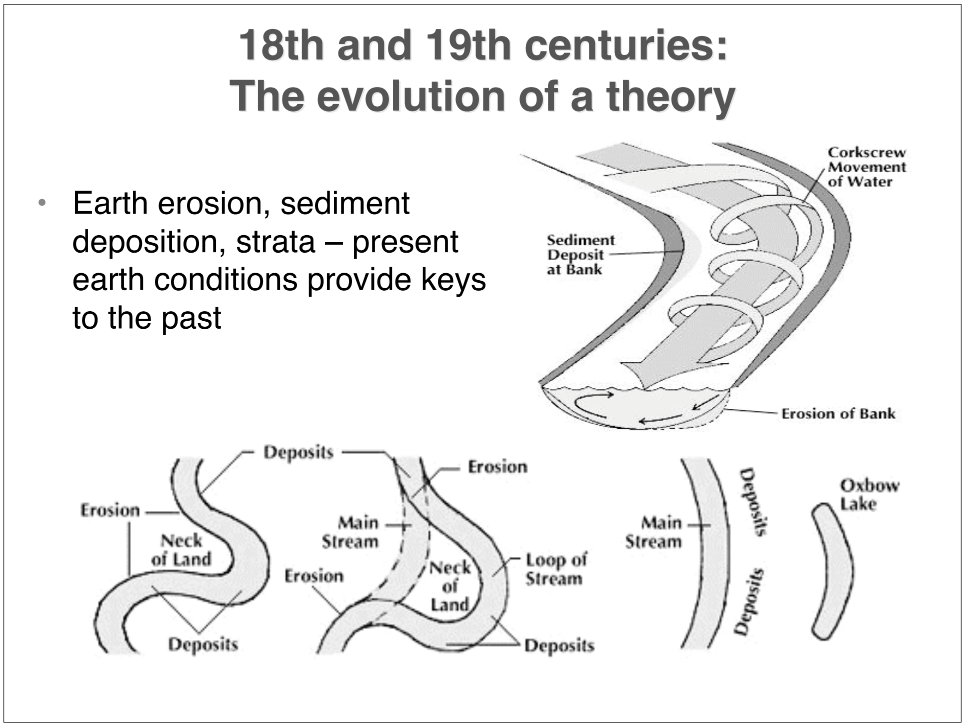
Why care about Evolutionary Analysis?

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

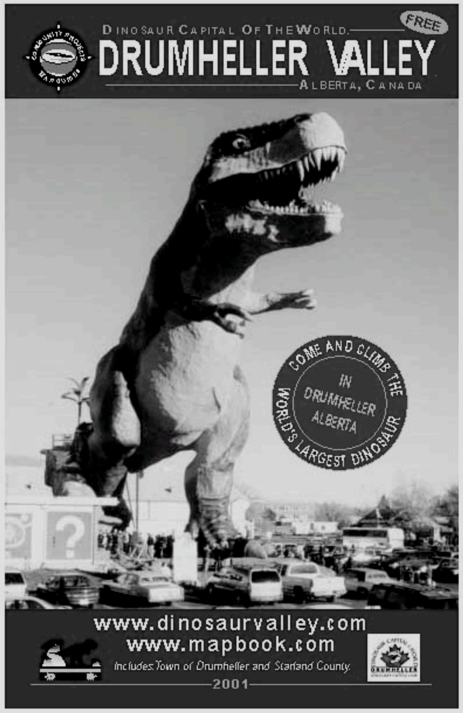


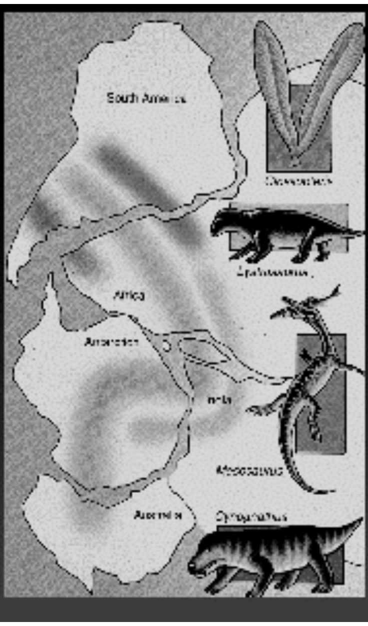
18th and 19th centuries: The evolution of a theory

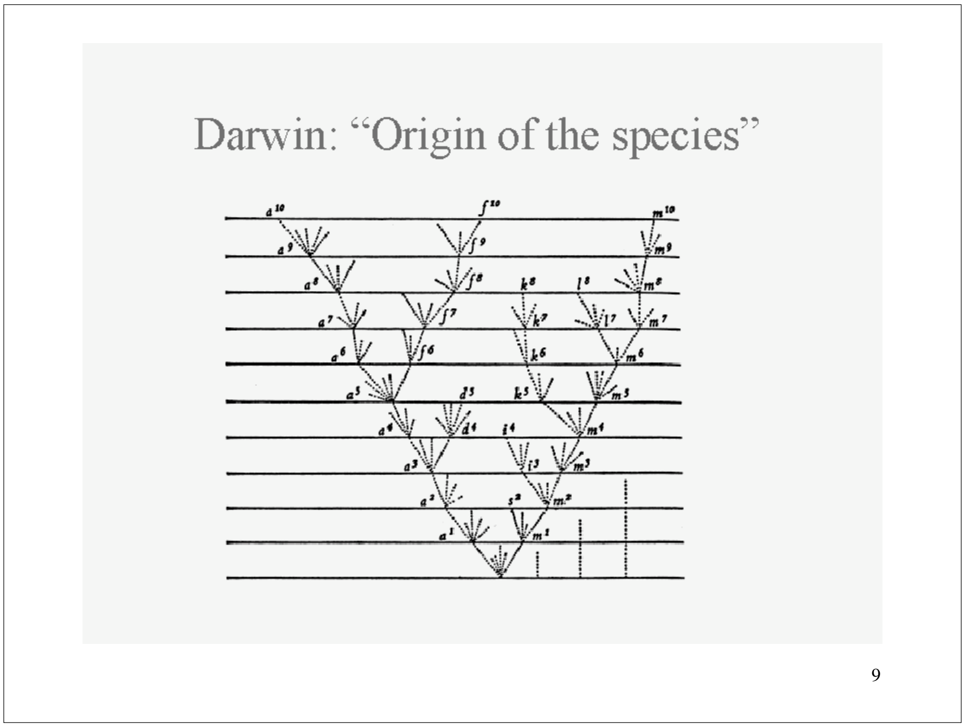
- Earth erosion, sediment deposition, strata – present earth conditions provide keys to the past

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

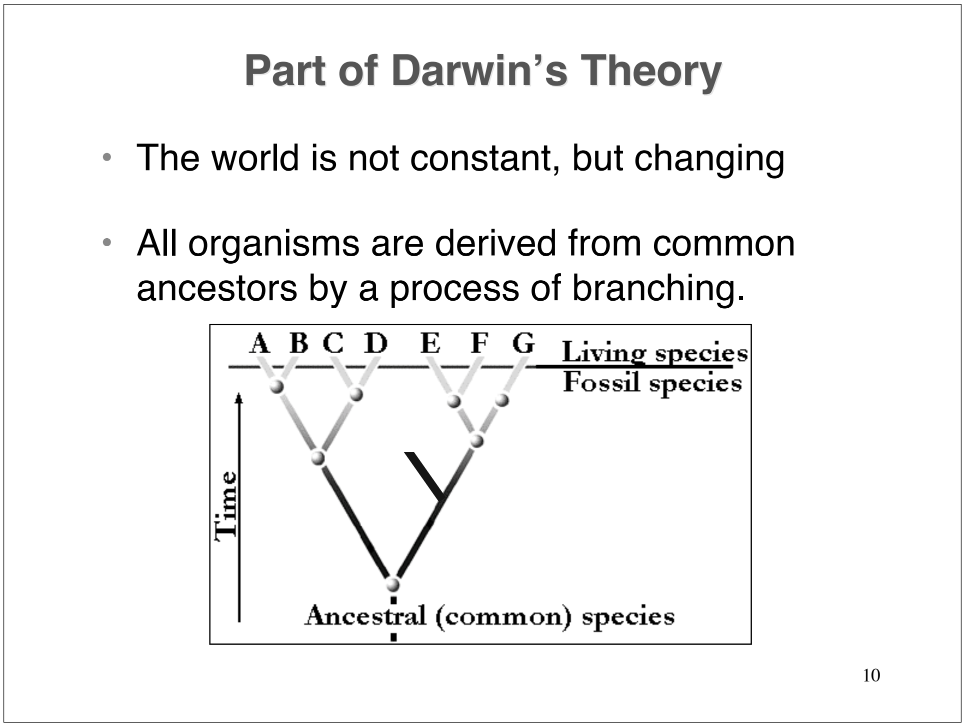

DINOSAUR CAPITAL OF THE WORLD
FREE
DRUMHELLER VALLEY
ALBERTA, CANADA
COME AND CLIMB THE WORLD'S LARGEST DINOSAUR
IN DRUMHELLER ALBERTA
www.dinosaurvalley.com
www.mapbook.com
Includes Town of Drumheller and Starland County
2001





Darwin: “Origin of the species”

9



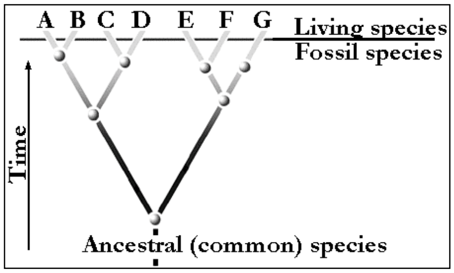
Part of Darwin’s Theory

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

10

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

11

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!

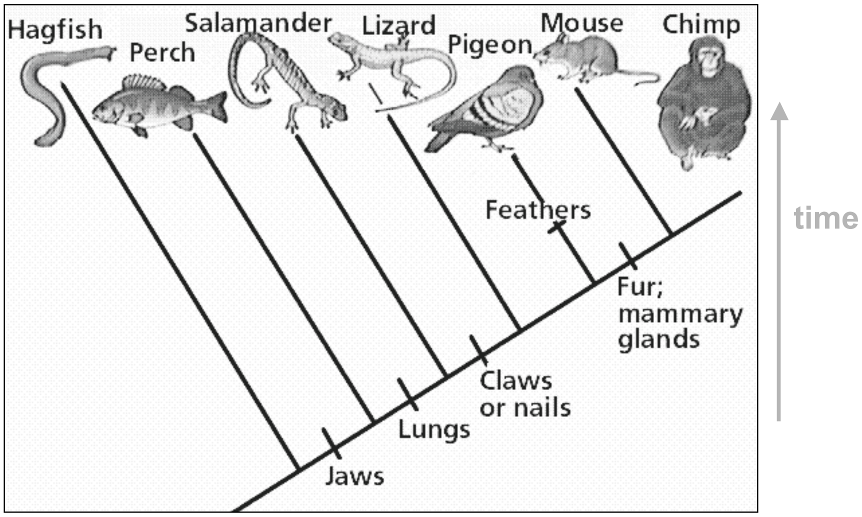
12

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

13

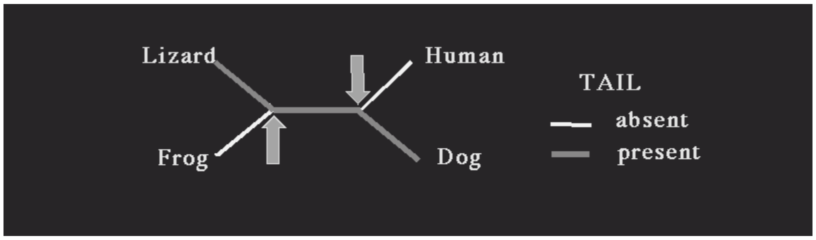
Evolution and unique characters



14

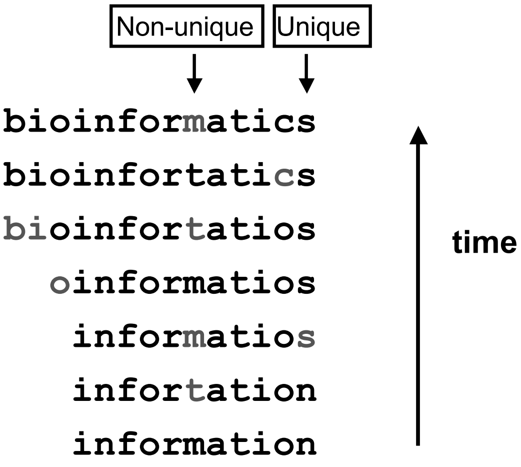
Homoplasy: The formation of tails

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

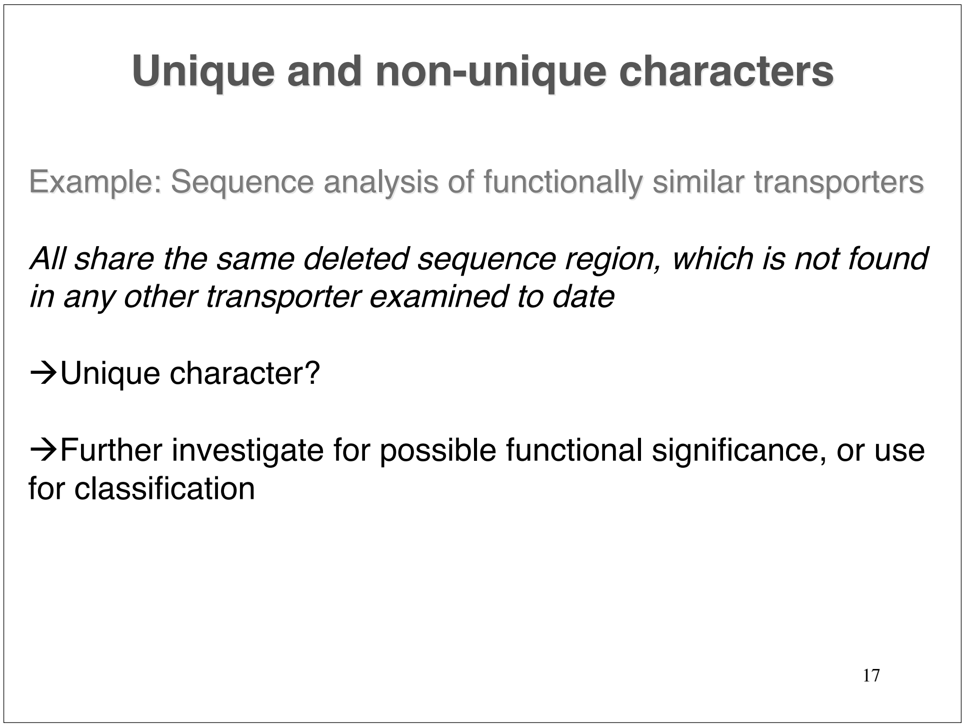


15

Unique and non-unique characters



16



Unique and non-unique characters

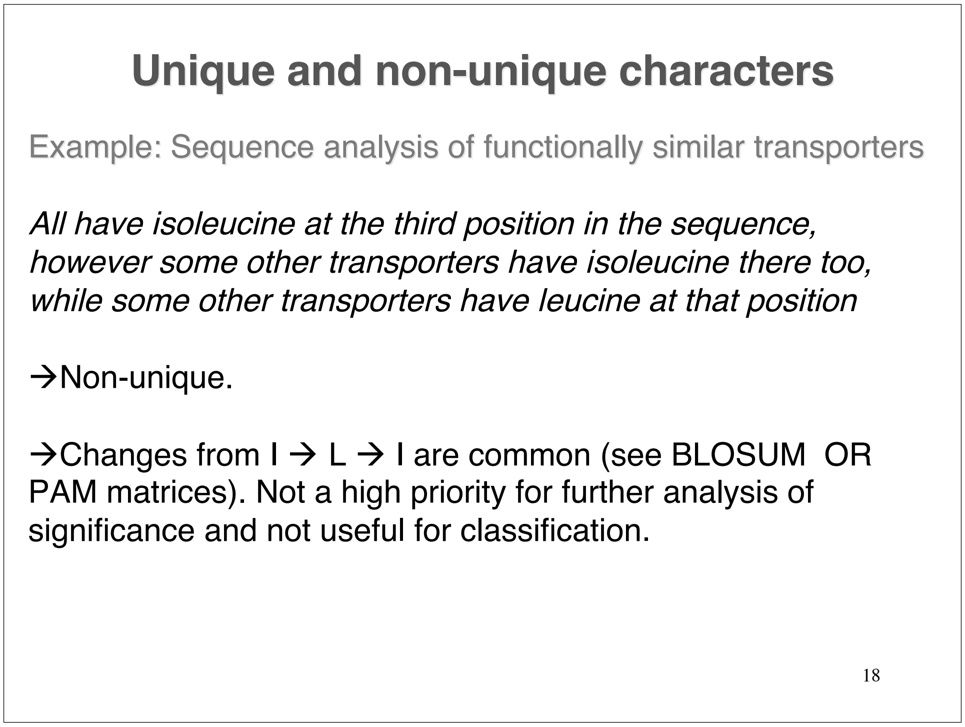
Example: Sequence analysis of functionally similar transporters

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

17



Unique and non-unique characters

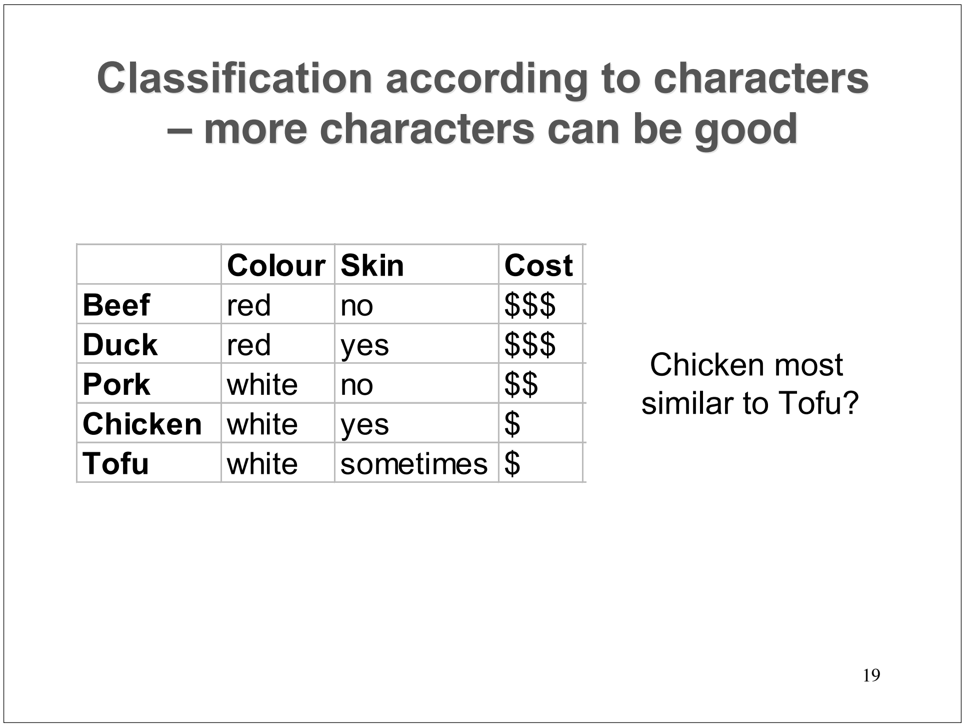
Example: Sequence analysis of functionally similar transporters

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.

18

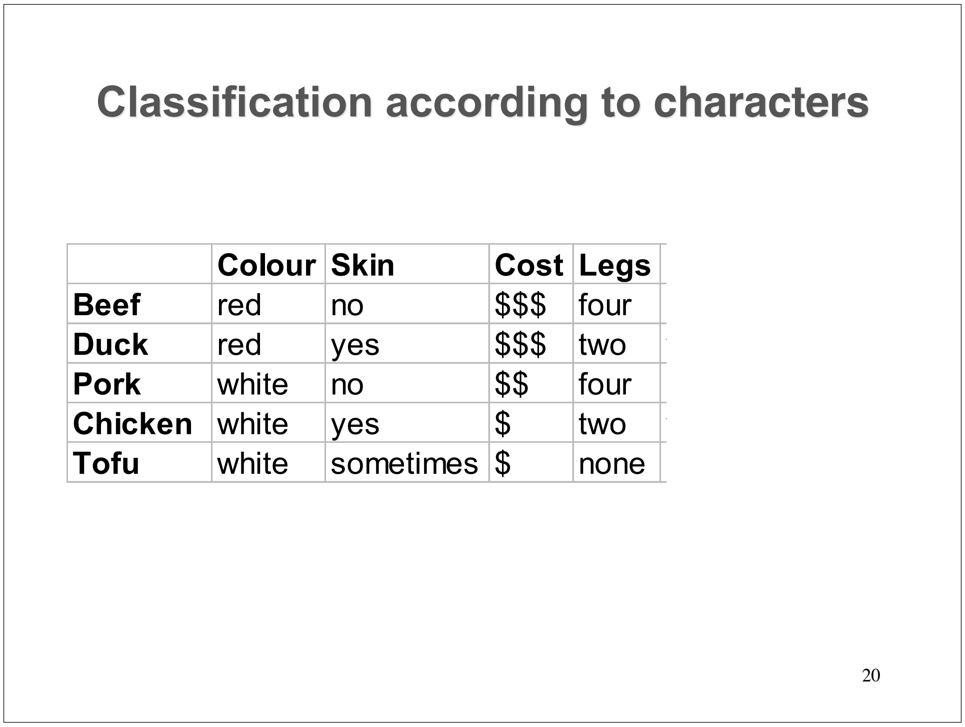


Classification according to characters – more characters can be good

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

Chicken most similar to Tofu?

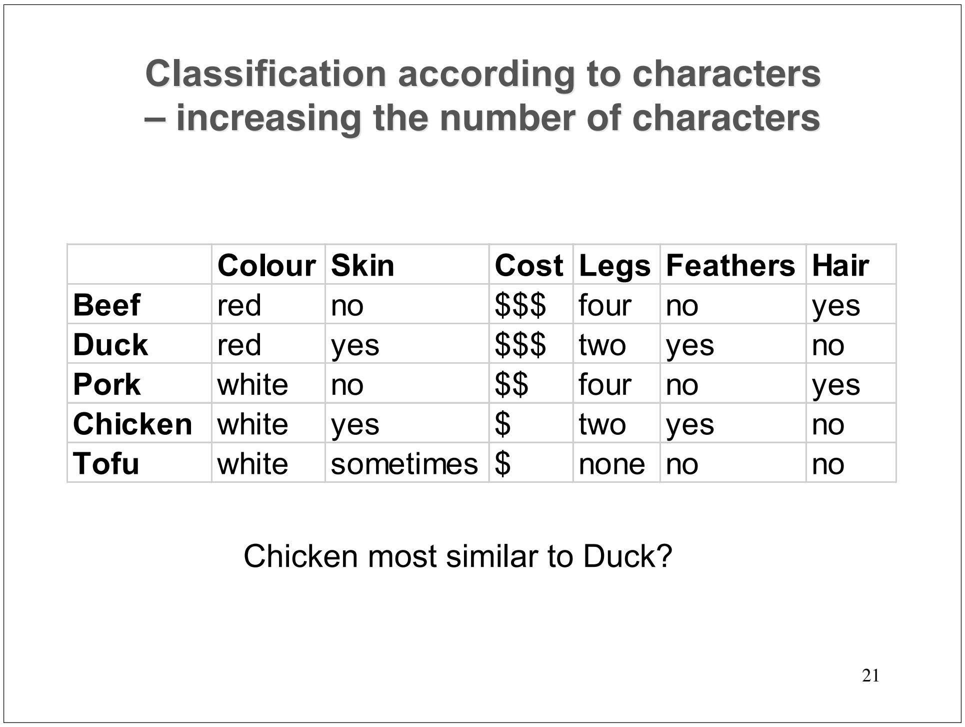
19



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

20

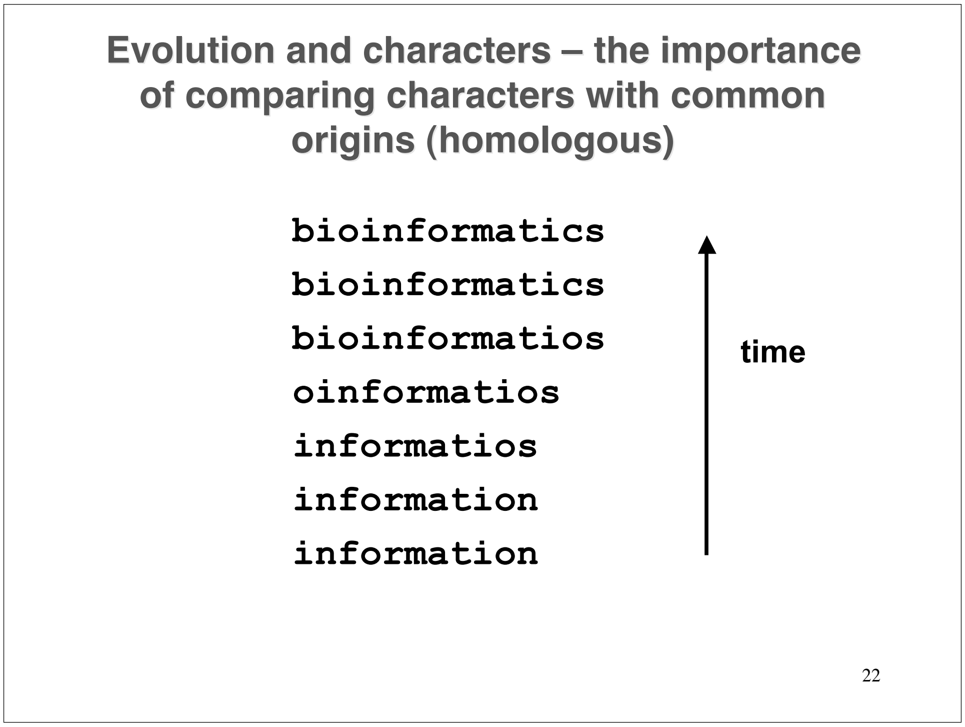


Classification according to characters – increasing the number of characters

	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?

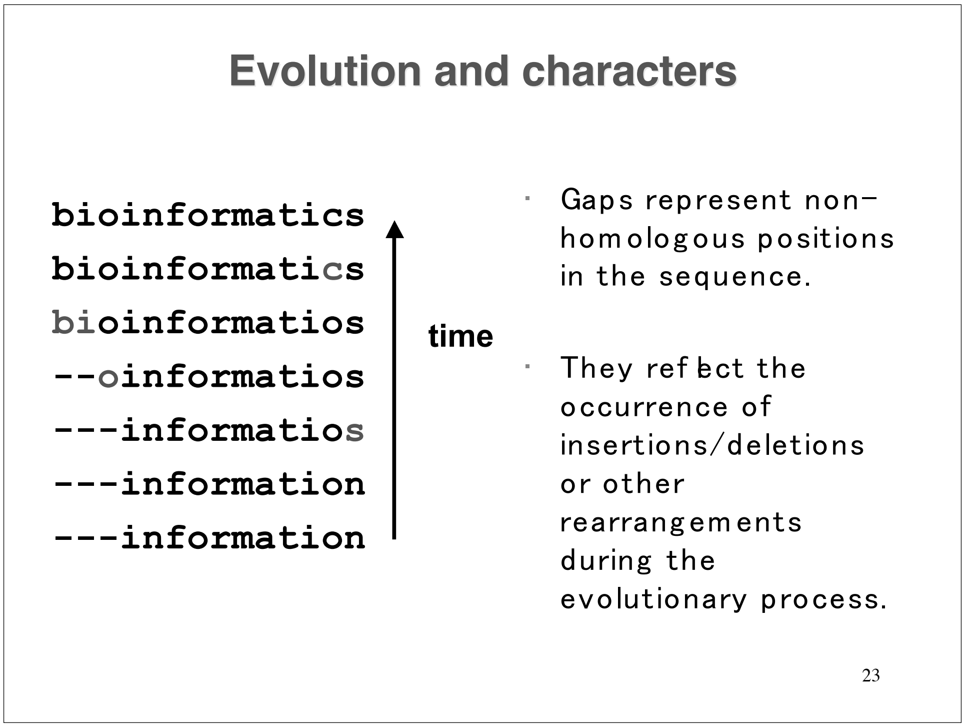
21



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

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

22

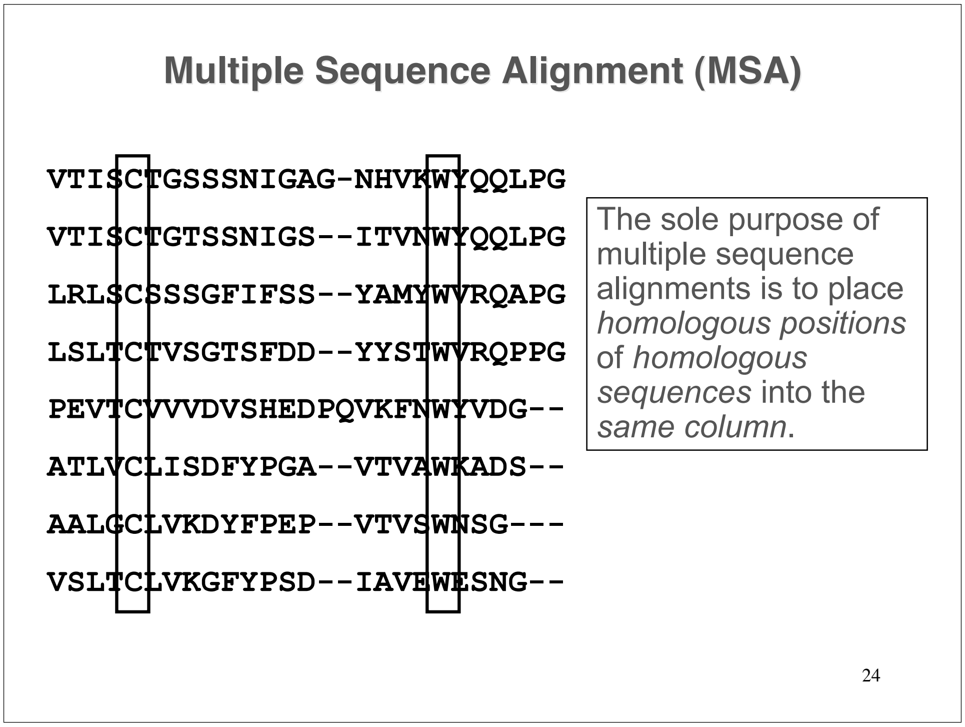


Evolution and characters

```
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.

23



Multiple Sequence Alignment (MSA)

```
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*.

24

Multiple sequence alignments (MSAs) and phylogenetic analysis

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

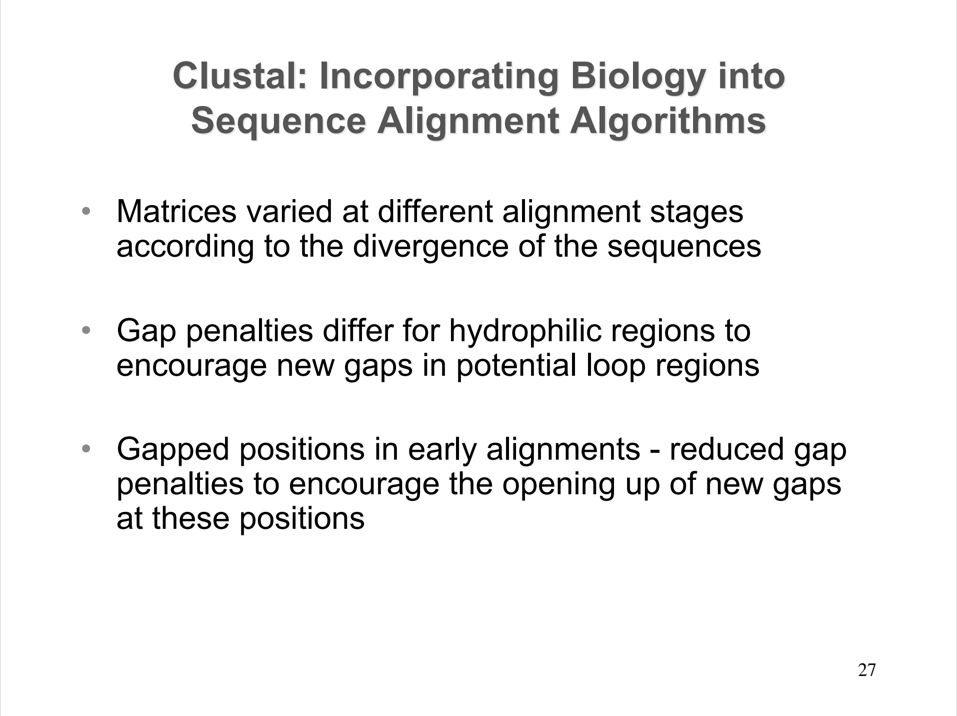
in → out!

25

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.

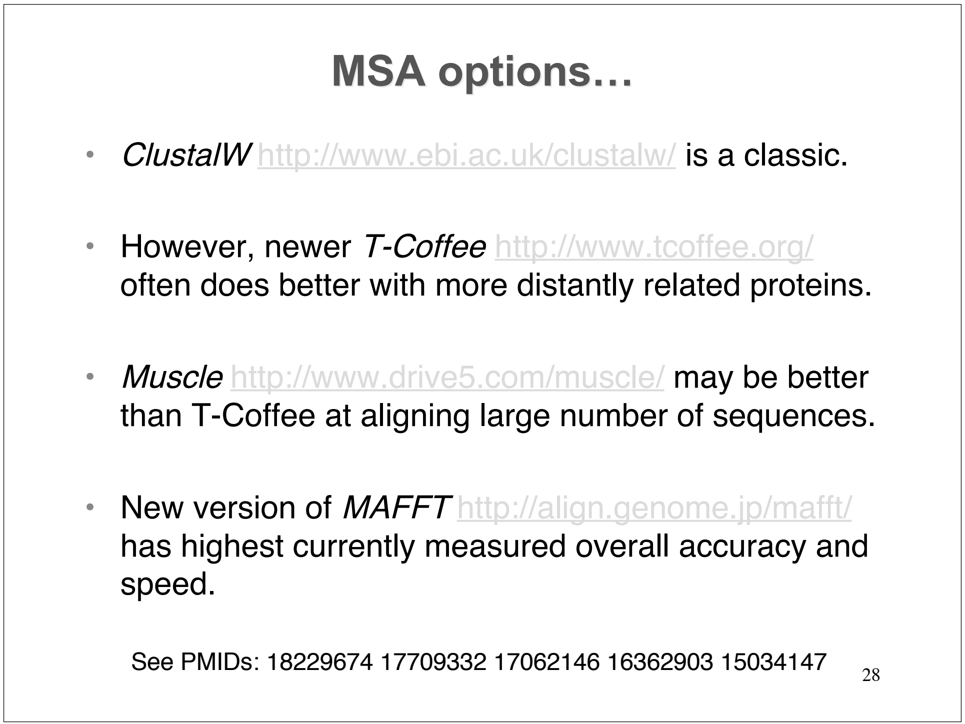
26



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

27



MSA options…

- *ClustalW* http://www.ebi.ac.uk/clustalw/ is a classic.
- However, newer *T-Coffee* http://www.tcoffee.org/ often does better with more distantly related proteins.
- *Muscle* http://www.drive5.com/muscle/ may be better than T-Coffee at aligning large number of sequences.
- New version of *MAFFT* http://align.genome.jp/mafft/ has highest currently measured overall accuracy and speed.

See PMIDs: 18229674 17709332 17062146 16362903 15034147

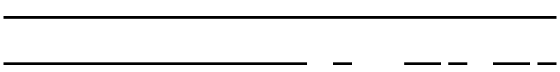
28

Standard MSA approach (first step for phylogenetic analysis)

- Be as sure as possible that the seq's included are homologous and avoid seq's with really different lengths
- Know as much as possible about the gene/protein in question before trying to create an alignment (secondary structure, domain structure…)
- Start with an automated alignment.
- If performing a fully automated procedure, consider using multiple accurate methods and compare where alignment differences occur. PMID: 17709332)

- If you can use a semi-manual approach, examine alignment:
  - Are you confident that aligned residues/bases evolved from a common ancestor?
  - Do the domains of the proteins/predicted secondary structures, etc. appear to be aligning correctly?

→ No? May 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..)

- Consider…. removing unreliably aligned regions for phylogenetic analysis

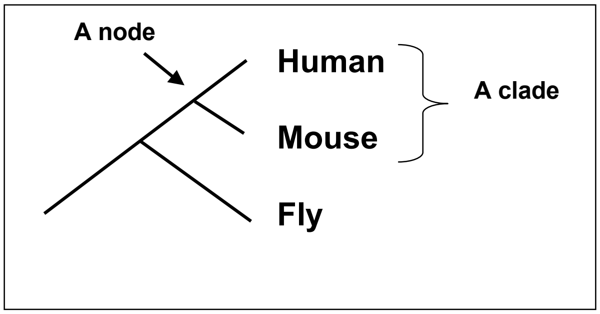
```
ILPITSPSKEGYESGKAPDEFSSGG
ILPEH--IKDDGELGAAPHSFSTAG
VLPLD-----S--AGRPADSFSAAG
VLPVDR-------DGQARDEYT-VG
VLPVDN-------KGEARDEYT-VG
LLPYDD-------QGRPQDDYSRAG
GIVSRSG---SNFDGEPKDSYGKVG
```

Delete?

- Move on to the tree construction stage of phylogenetic analysis!

A Douglas Fir in Vancouver, BC

A phylogenetic tree

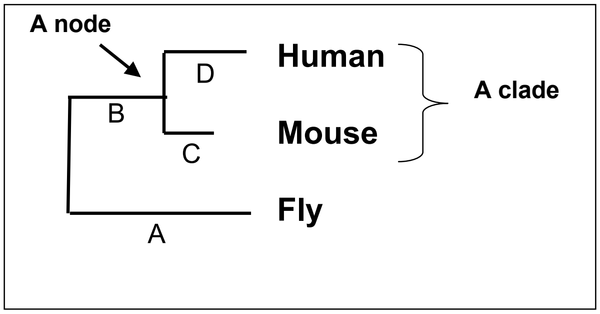


taxon -- Any named group of organisms – evolutionary theory not necessarily involved.

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

33

A phylogenetic tree with branch lengths

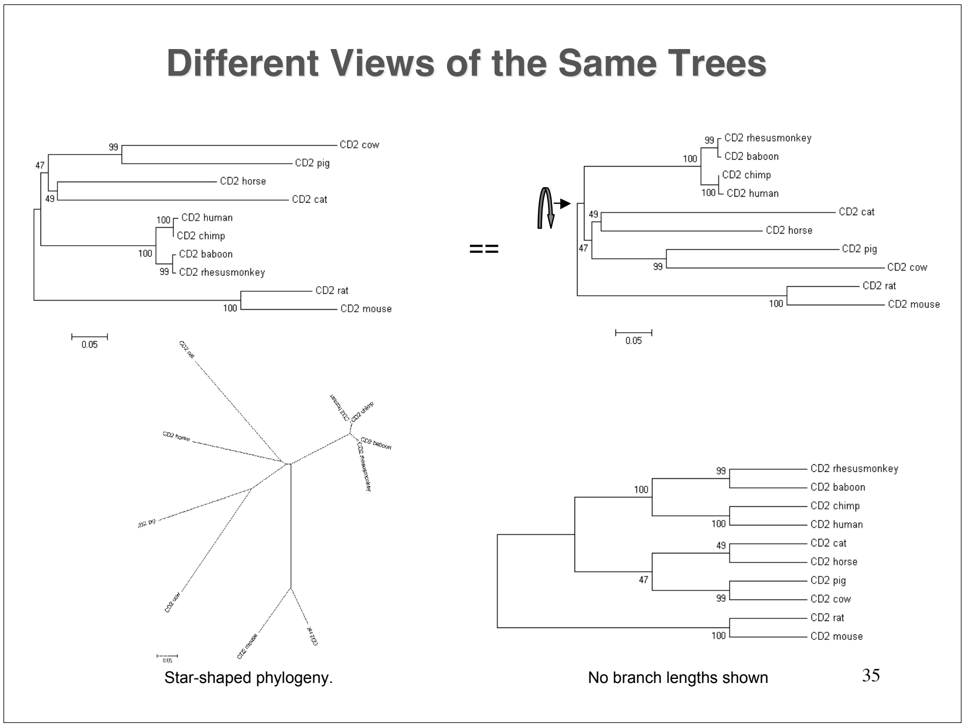


Branch length can be significant…

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)

34



Different Views of the Same Trees

Star-shaped phylogeny.

No branch lengths shown

35

Phylogenetic analysis

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

36

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

↓

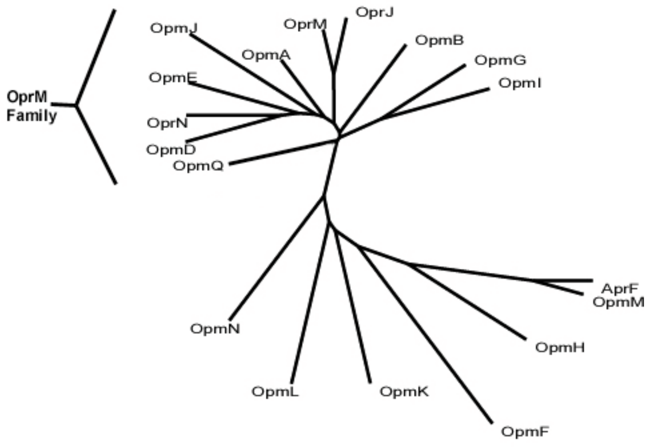
“Whole genome” datasets of genes

↓

rRNA genes and multiple suitable genes

37

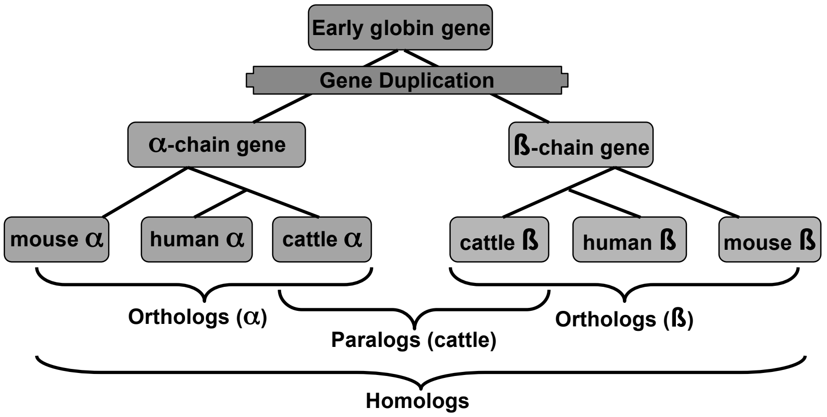
Gene/Protein Relationships



Homolog, ortholog, paralog??

38

Orthologous or paralogous homologs



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

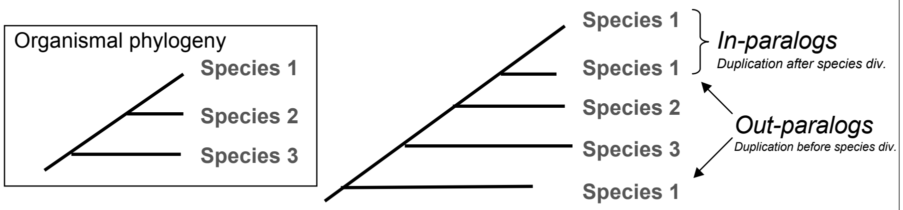
Paralogs – diverged after gene duplication – ***some functional divergence occurs***

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

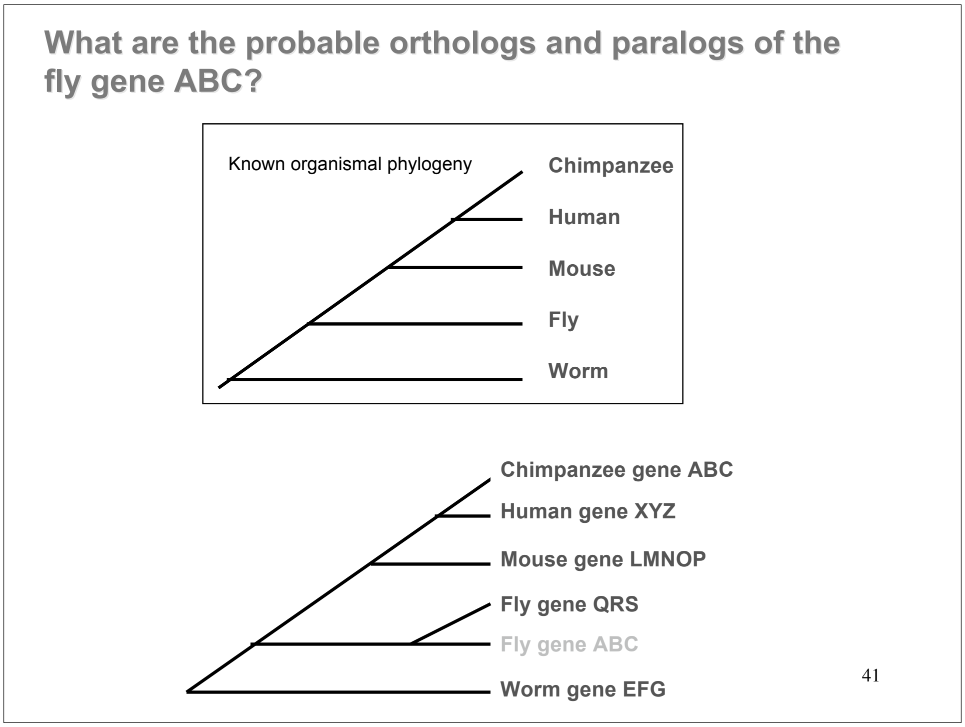
Identifying Gene/Protein Relationships from Phylogenetic trees

- **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 gene ABC?

Known organismal phylogeny

Chimpanzee

Human

Mouse

Fly

Worm

Chimpanzee gene ABC

Human gene XYZ

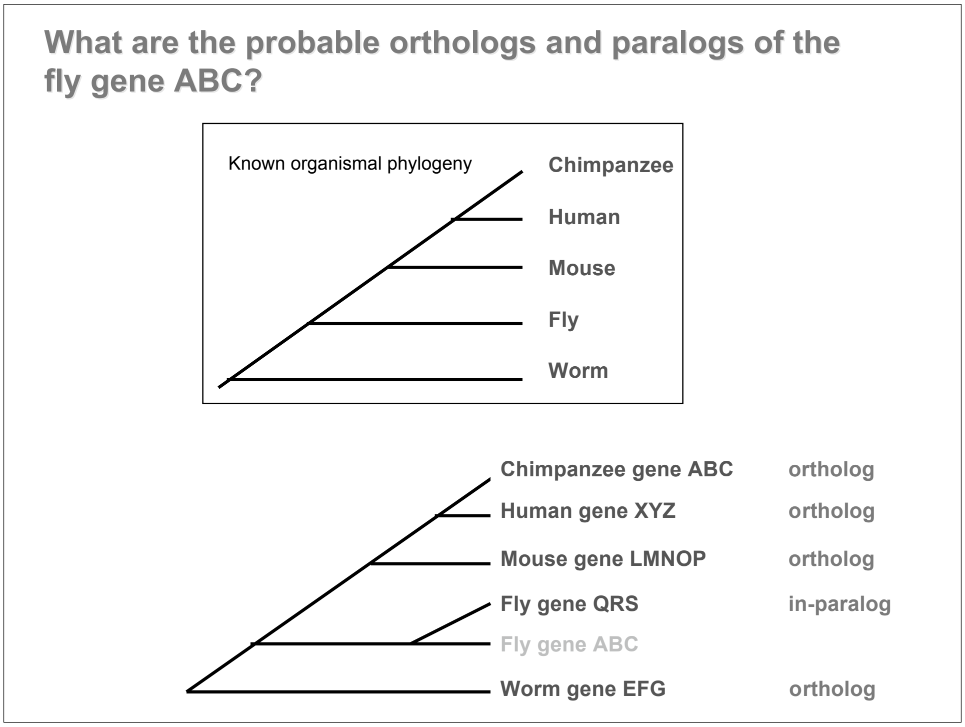
Mouse gene LMNOP

Fly gene QRS

Fly gene ABC

Worm gene EFG

41



What are the probable orthologs and paralogs of the fly gene ABC?

Known organismal phylogeny

Chimpanzee

Human

Mouse

Fly

Worm

Chimpanzee gene ABC ortholog

Human gene XYZ ortholog

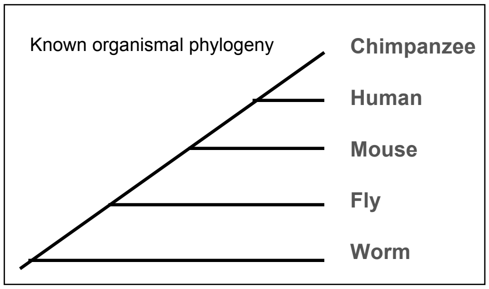
Mouse gene LMNOP ortholog

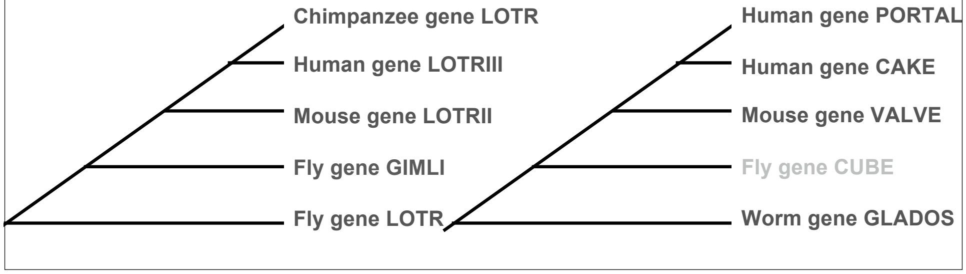
Fly gene QRS in-paralog

Fly gene ABC

Worm gene EFG ortholog

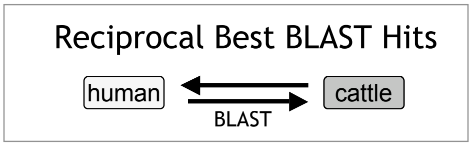
What are the probable orthologs and paralogs of the fly genes GIMLI and CUBE?





High Throughput Gene Orthology: How to detect?

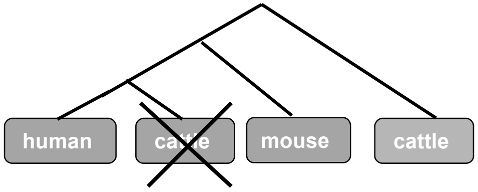
- Most common high throughput computational method: Identify reciprocal best BLAST hits (EGO, COGs,…) and cluster (INPARANOID…)



Example Problem:

- If making comparisons between human and bovine, for example, the bovine gene dataset is still quite incomplete

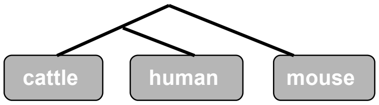
- Therefore, current best hit may be a paralog now and the true ortholog not yet sequenced



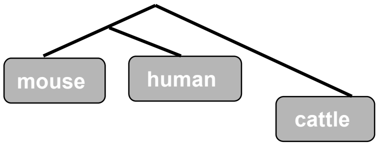
Can we improve orthology analysis for linking functionally similar genes?

- One solution: Phylogenetic analysis of all putative orthologs, using one or more species as an outgroup
- Assumption for the case below:
  - Mouse and Human gene datasets are more complete, with more true orthologs identified

Expect (organismal phylogeny):

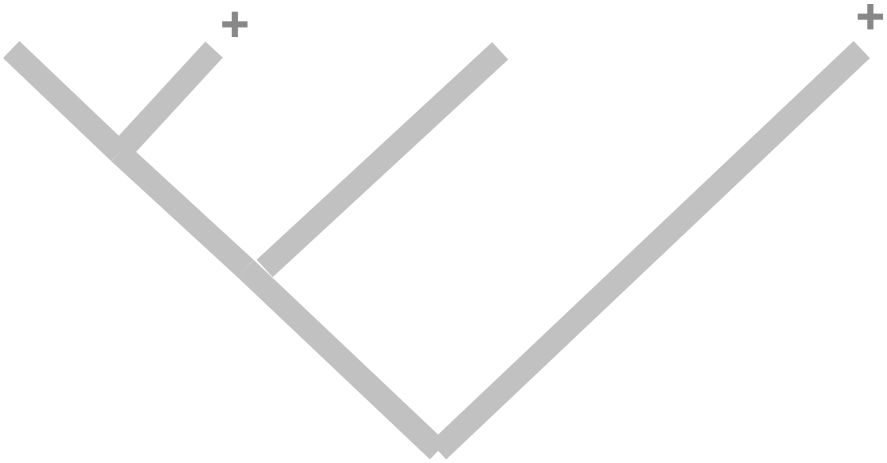


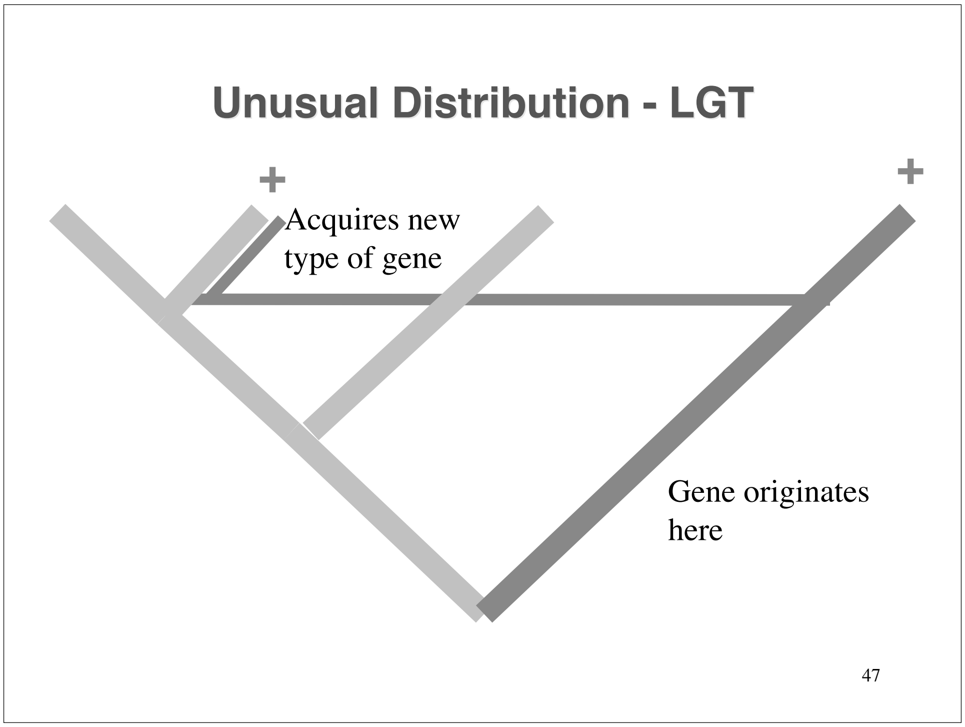
Reject:



Ortholuge software: The beginnings of an automated attempt finds that 1 in 20 genes have unusual divergence! PMID: 16729895

Unusual gene distribution pattern in species? …remember all the possible explanations!



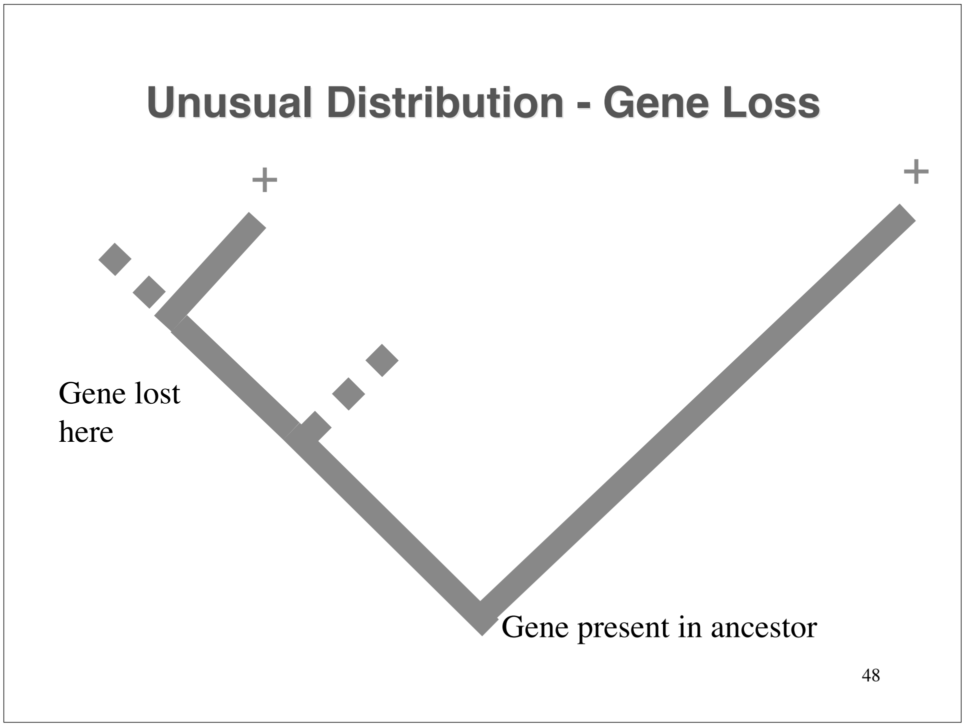


Unusual Distribution - LGT

Acquires new type of gene

Gene originates here

47



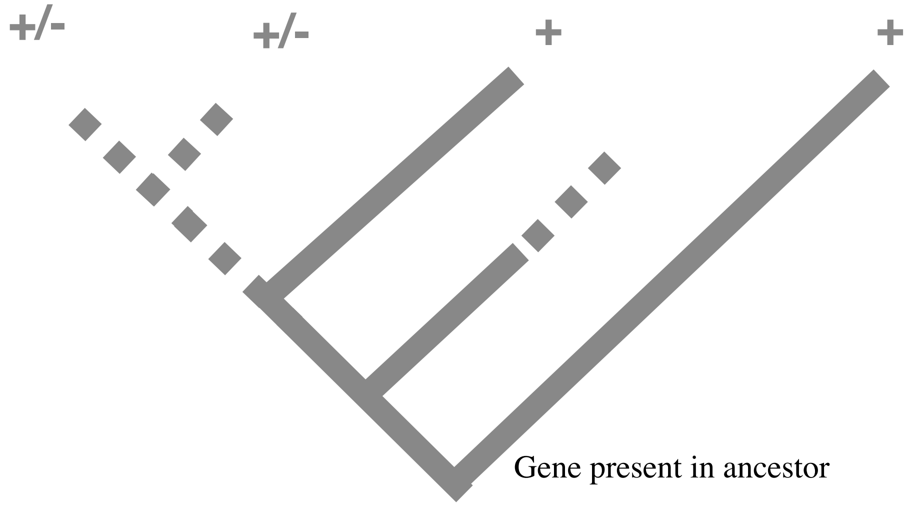
Unusual Distribution - Gene Loss

Gene lost here

Gene present in ancestor

48

Unusual Distribution - Incomplete Data

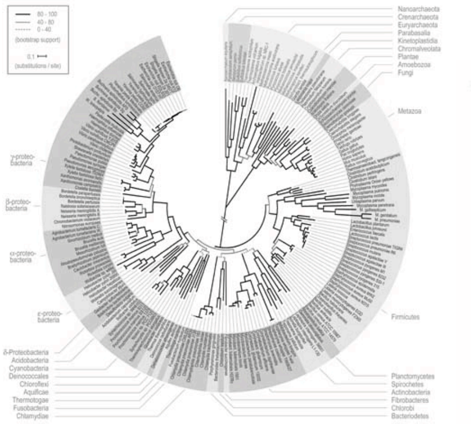


Gene present in ancestor

49

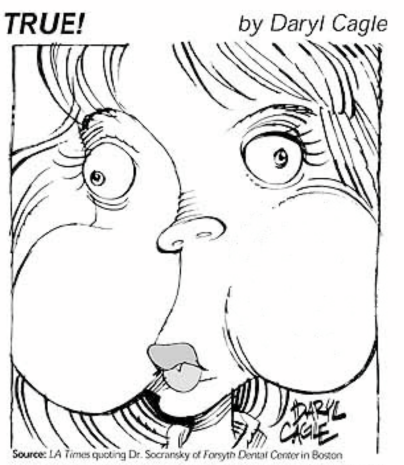
Hope for the future

Better sampling of all the species in our world



2004: Environmental genomics sampling takes centre stage

Tyson et al (2004) *Nature*, 428, 37-43.
Venter et al (2004) *Science,* 304, 66-74.

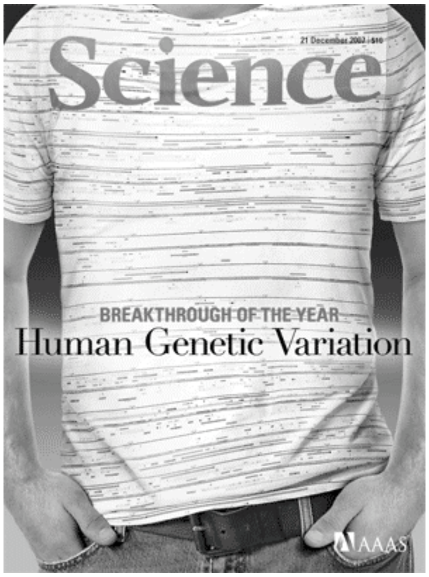


Source: *LA Times quoting Dr. Socransky of Forsyth Dental Center in Boston*

The number of bacteria living in your mouth can easily exceed the number of people who live on the Earth.

Hope for the future

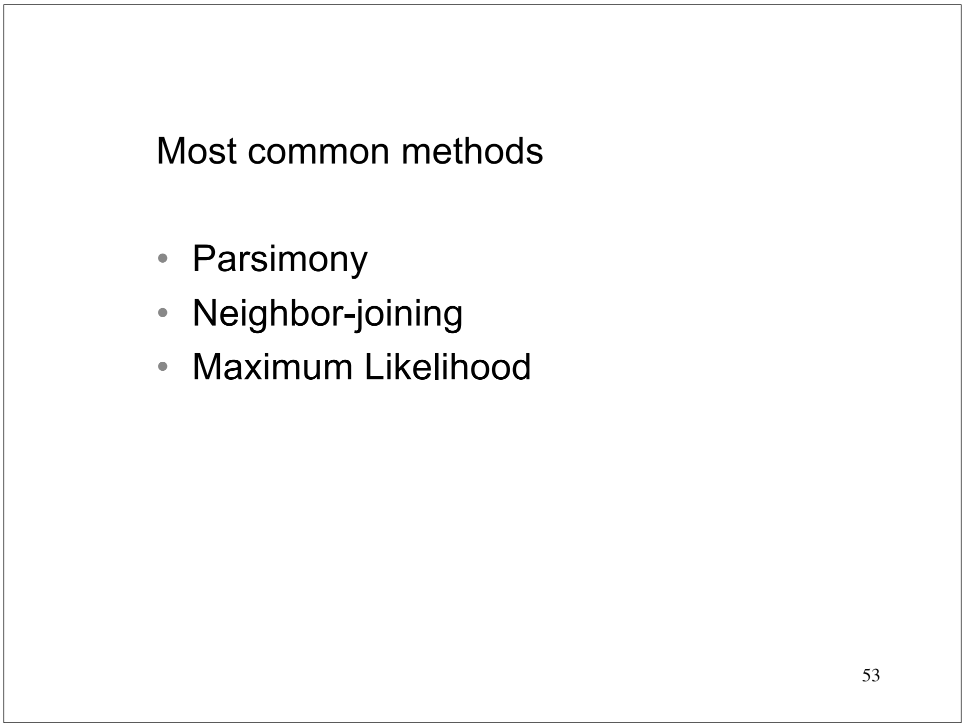
Better sampling <u>within</u> species in our world



51

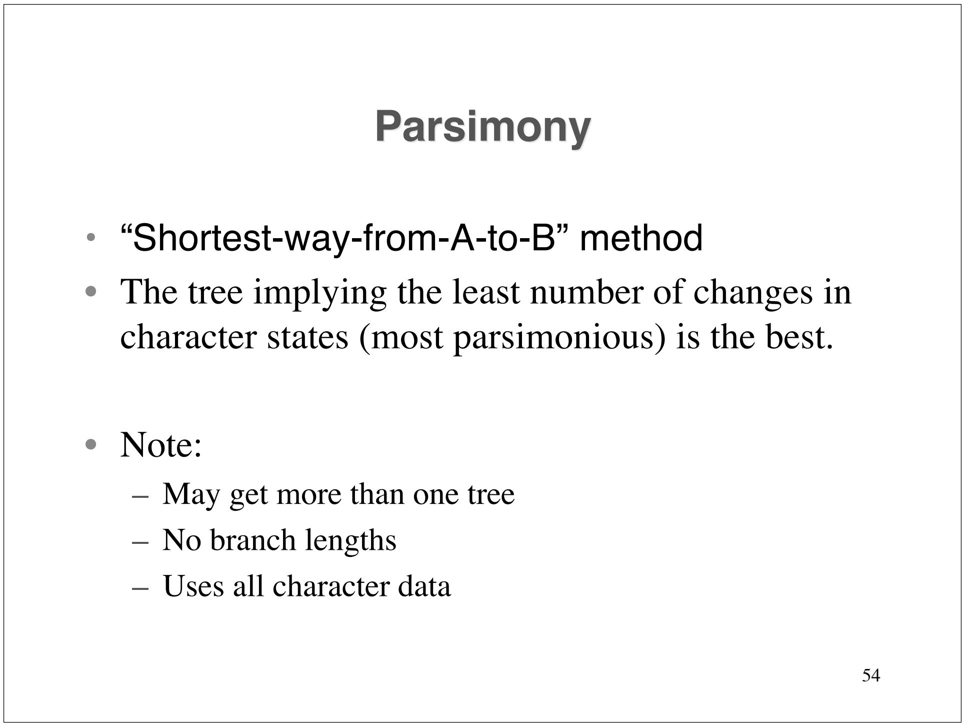
"So….. how do we construct a phylogenetic tree??"

52



Most common methods

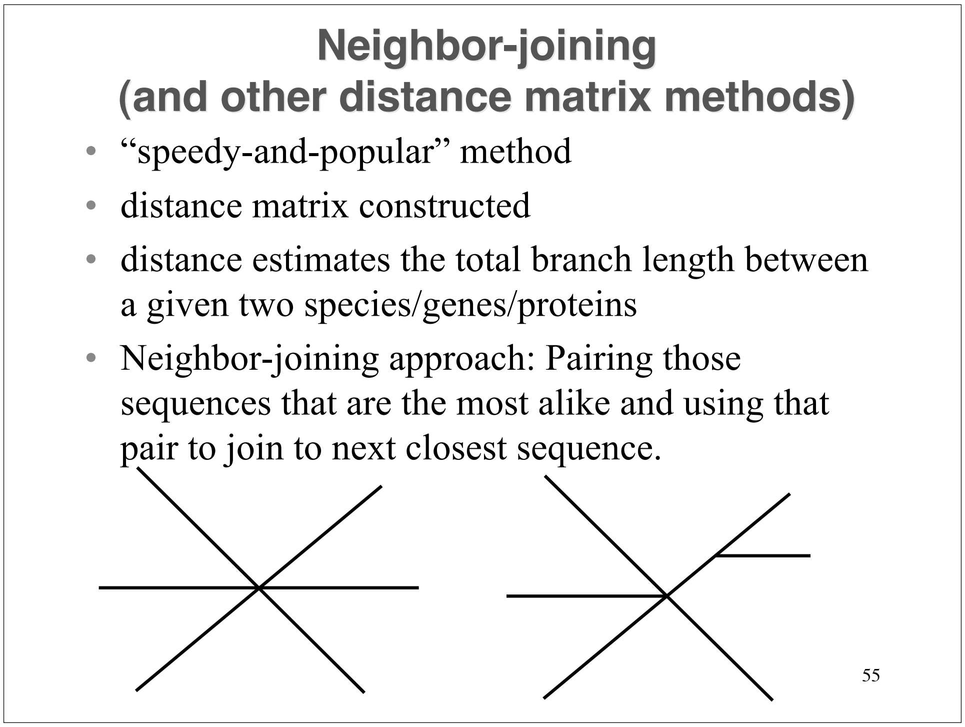
- Parsimony
- Neighbor-joining
- Maximum Likelihood



Parsimony

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

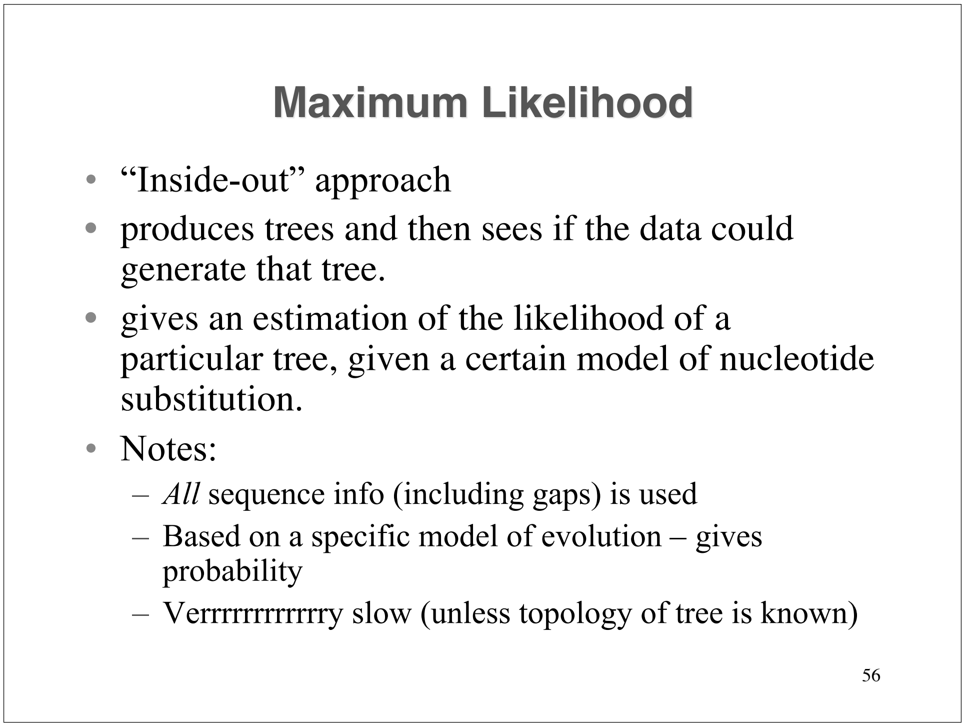
- Note:
  - May get more than one tree
  - No branch lengths
  - Uses all character data



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.

55



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)

56

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

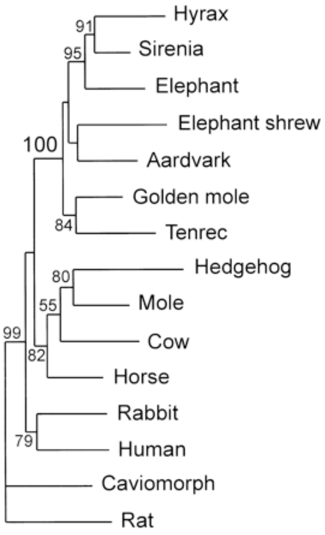
57

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

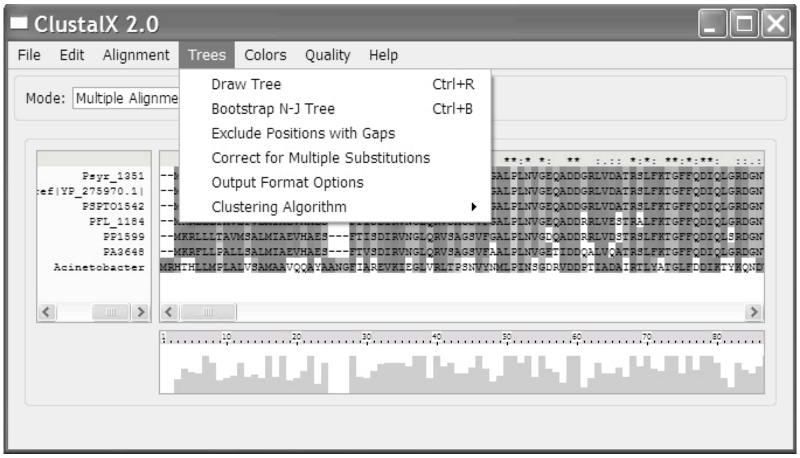


58

Phylogenetic Tree Construction: Examples of Common Software

ClustalX *http://bips.u-strasbg.fr/fr/Documentation/ClustalX/*

- **Incorporates a simple neighbor-joining method with MSA**
- **Good for a quick view of the phylogeny**



```
                                                    **:* *:  **  :.:: *:*: **:*:**:  :::
      Psyr_1351 --M                         GALPLNVGEQADDGRLVDATRSLFKTGFFQDIQLGRDGN
ef|YP_275970.1| --M                         GALPLNVGEQADDGRLVDATRSLFKTGFFQDIQLGRDGN
      PSPTO1542 --M                         GALPLNVGEQADDGRLVDATRSLFKTGFFQDIQLGRDGN
       PFL_1184 --                           GALPLNVGEQADDRRLVESTRALFKTGFFQDIQLGRDGN
         PP1599 --MKRLLLTAVMSALMIAEVHAES---FTISDIRVNGLQRVSAGSVFGALPLNVGDQADDRRLVDSTRSLFKTGFFQDIQLSRDGN
         PA3648 --MKRFLLPALLSALMIAEVHAES---FTVSDIRVNGLQRVSAGSVFAALPLNVGETIDDQALVQATRSLFKTGFFQDIQLGRDGN
  Acinetobacter MRHTHLLMPLALVSAMAAVQQAYAANGFIAREVKIEGLVRLTPSNVYNMLPINSGDRVDDPTIADAIRTLYATGLFDDIKTYKQND
```

Phylogenetic Tree Construction: Examples of Common Software

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

ClustalX (incorporates a simple neighbor joining method with MSA – good for a quick view of the phylogeny)
http://bips.u-strasbg.fr/fr/Documentation/ClustalX/

PHYLIP (a classic – many web-based versions also made)
http://evolution.genetics.washington.edu/phylip.html

PAUP
http://paup.csit.fsu.edu/

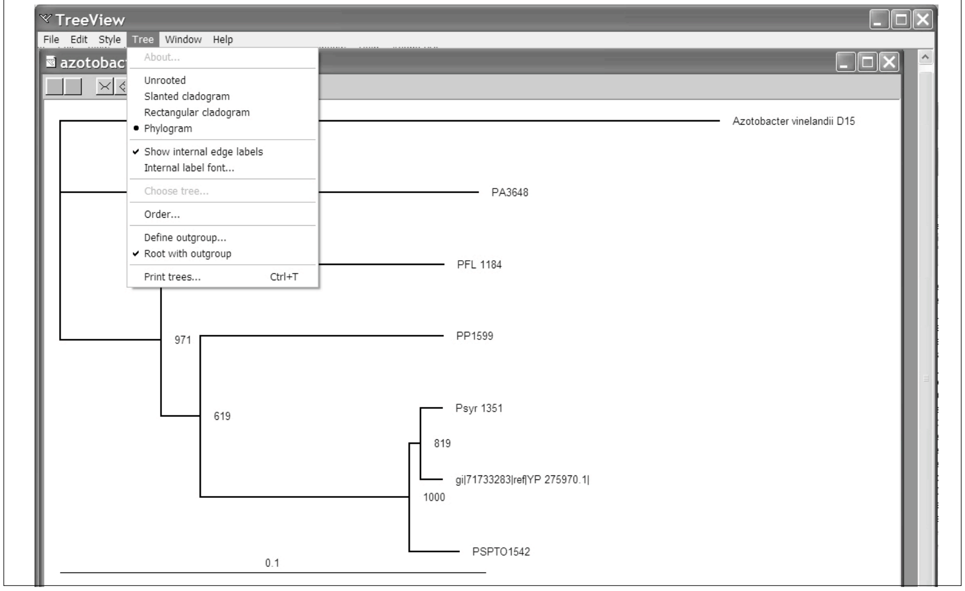
MEGA 2.1
www.megasoftware.net/

60

Phylogenetic Tree Viewing

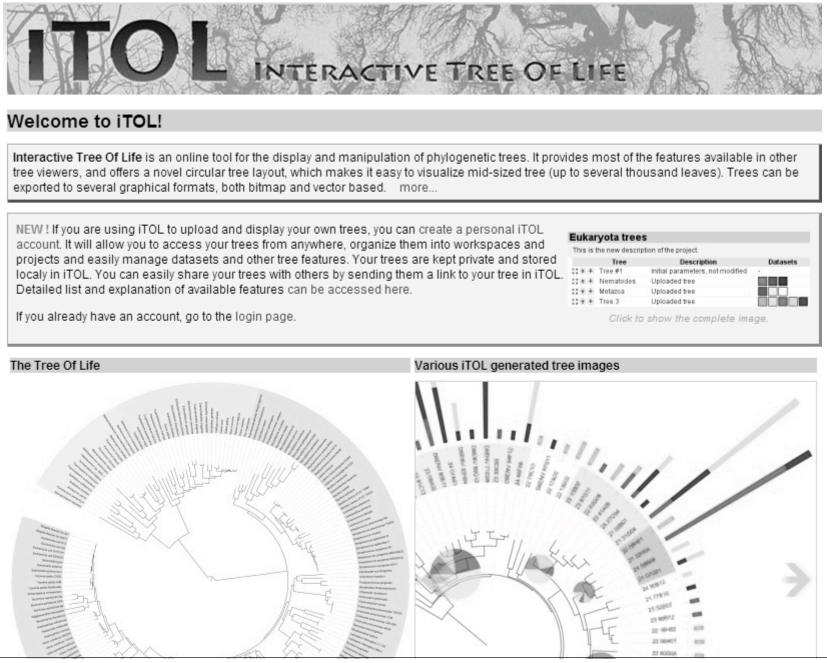
TREEVIEW

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



Phylogenetic Tree Viewing

http://itol.embl.de/



Welcome to iTOL!

Interactive Tree Of Life is an online tool for the display and manipulation of phylogenetic trees. It provides most of the features available in other tree viewers, and offers a novel circular tree layout, which makes it easy to visualize mid-sized tree (up to several thousand leaves). Trees can be exported to several graphical formats, both bitmap and vector based. more...

NEW ! If you are using iTOL to upload and display your own trees, you can create a personal iTOL account. It will allow you to access your trees from anywhere, organize them into workspaces and projects and easily manage datasets and other tree features. Your trees are kept private and stored locally in iTOL. You can easily share your trees with others by sending them a link to your tree in iTOL. Detailed list and explanation of available features can be accessed here.

If you already have an account, go to the login page.

Eukaryota trees

This is the new description of the project.

Tree	Description	Datasets
Tree #1	Initial parameters, not modified	-
Nematodes	Uploaded tree	■■■
Metazoa	Uploaded tree	■□□
Tree 3	Uploaded tree	■□■□■

Click to show the complete image.

The Tree Of Life

Various iTOL generated tree images

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 out of date.

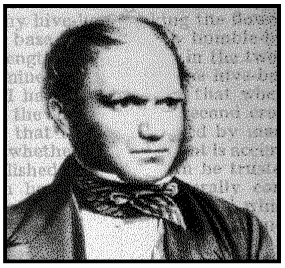
Nei, Masatoshi & Kumar, Sudhir. 2000. Molecular evolution and phylogenetics Oxford ; New York. Oxford University Press.

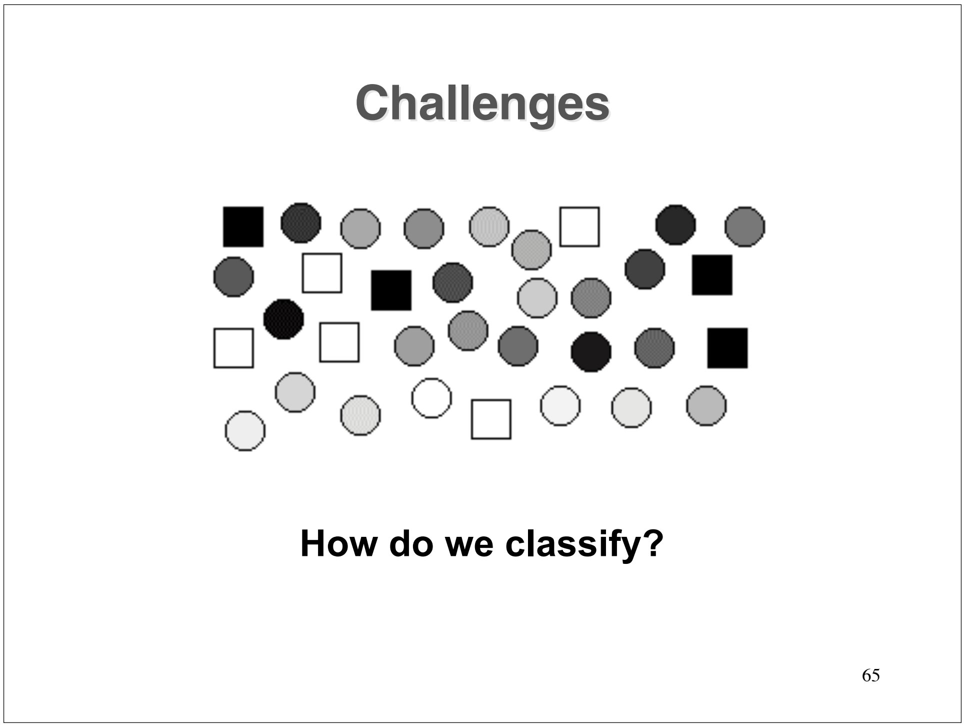
- by two very well respected researchers in the field. A bit more in-depth than the previous book, and generally very useful.

63

Evolutionary Analysis: Key Concepts

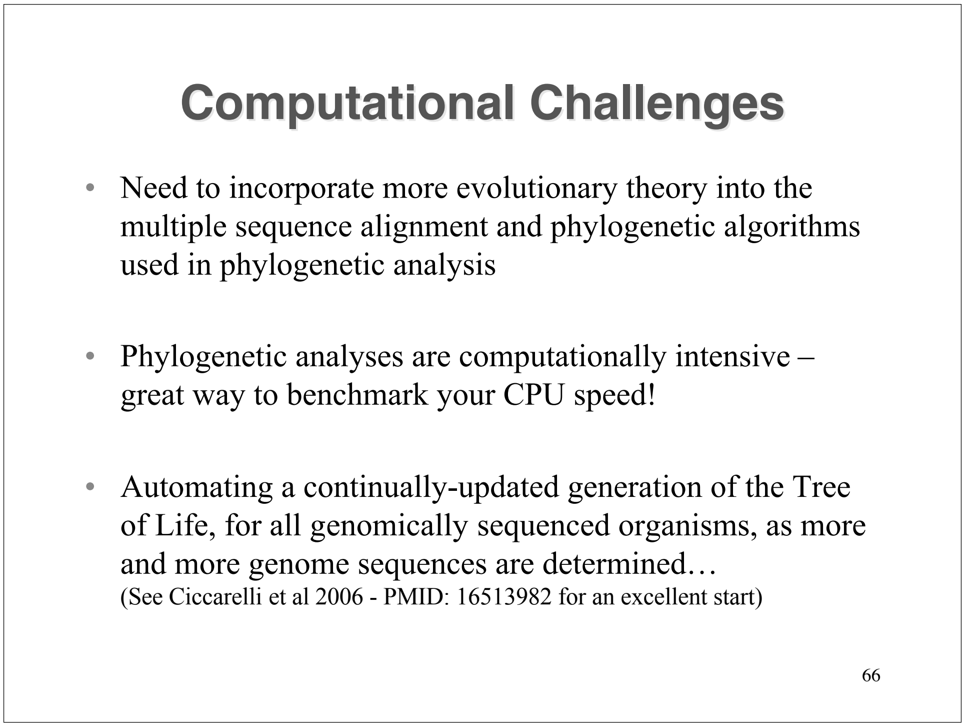
- Foundation of most bioinformatic analyses: Evolutionary theory
- Unique verses non-unique characters
- Sequence alignments are important!
- Fundamentals of phylogenetics and interpreting phylogenetic trees (with cautionary notes)
- Overview of some common phylogenetic methods
- Appreciate the need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





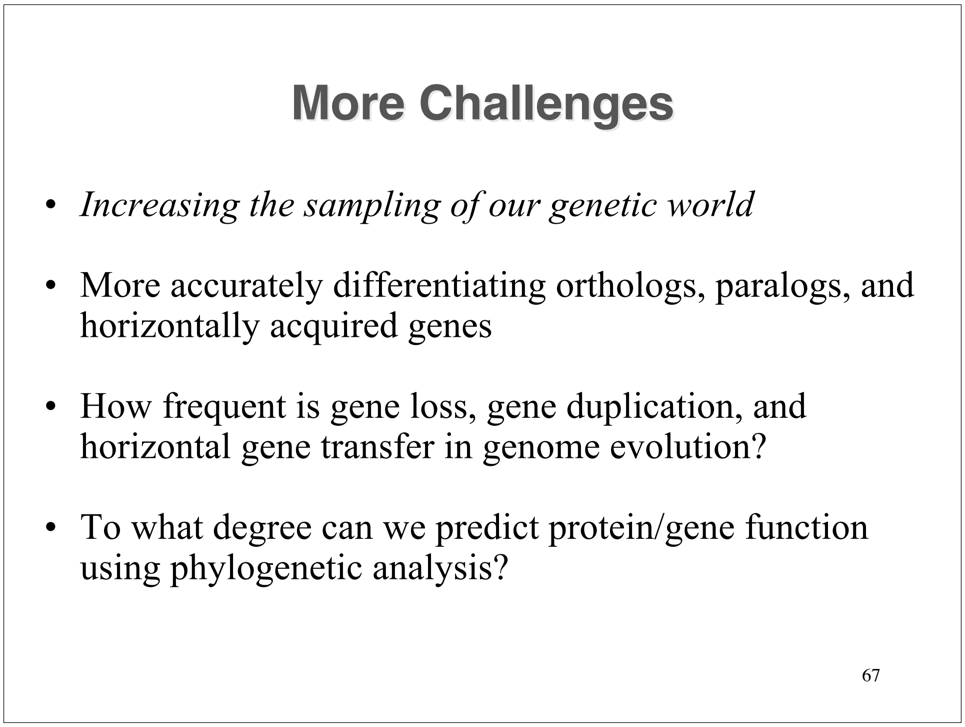
Challenges

How do we classify?



Computational Challenges

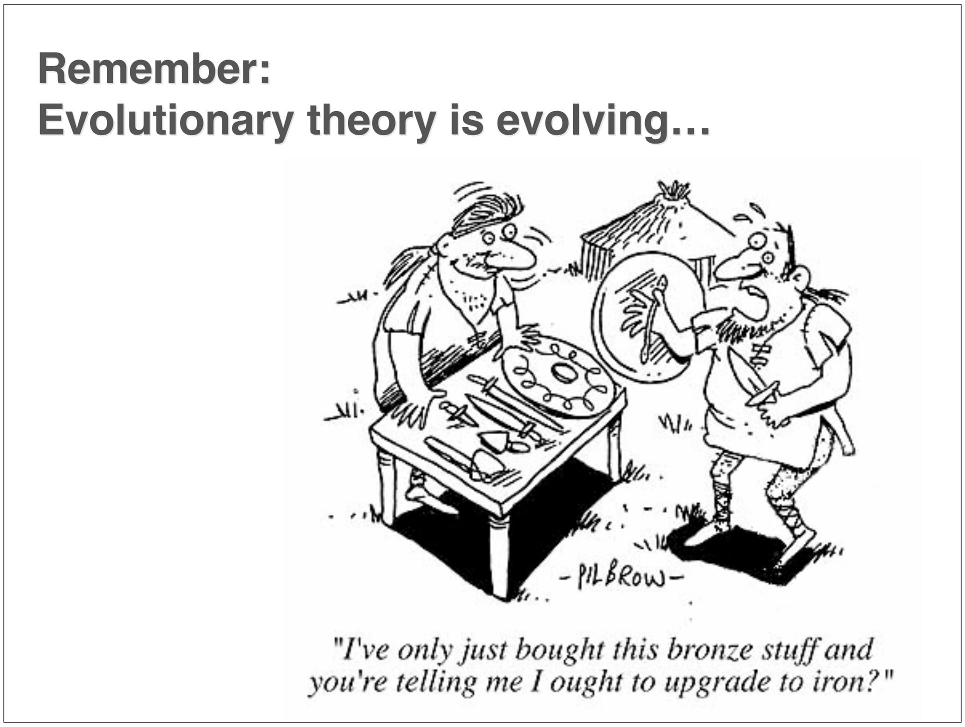
- Need to incorporate more evolutionary theory into the multiple sequence alignment and phylogenetic algorithms used in phylogenetic analysis
- Phylogenetic analyses are computationally intensive – great way to benchmark your CPU speed!
- Automating a continually-updated generation of the Tree of Life, for all genomically sequenced organisms, as more and more genome sequences are determined…
  (See Ciccarelli et al 2006 - PMID: 16513982 for an excellent start)



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?

67



Remember: Evolutionary theory is evolving…

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