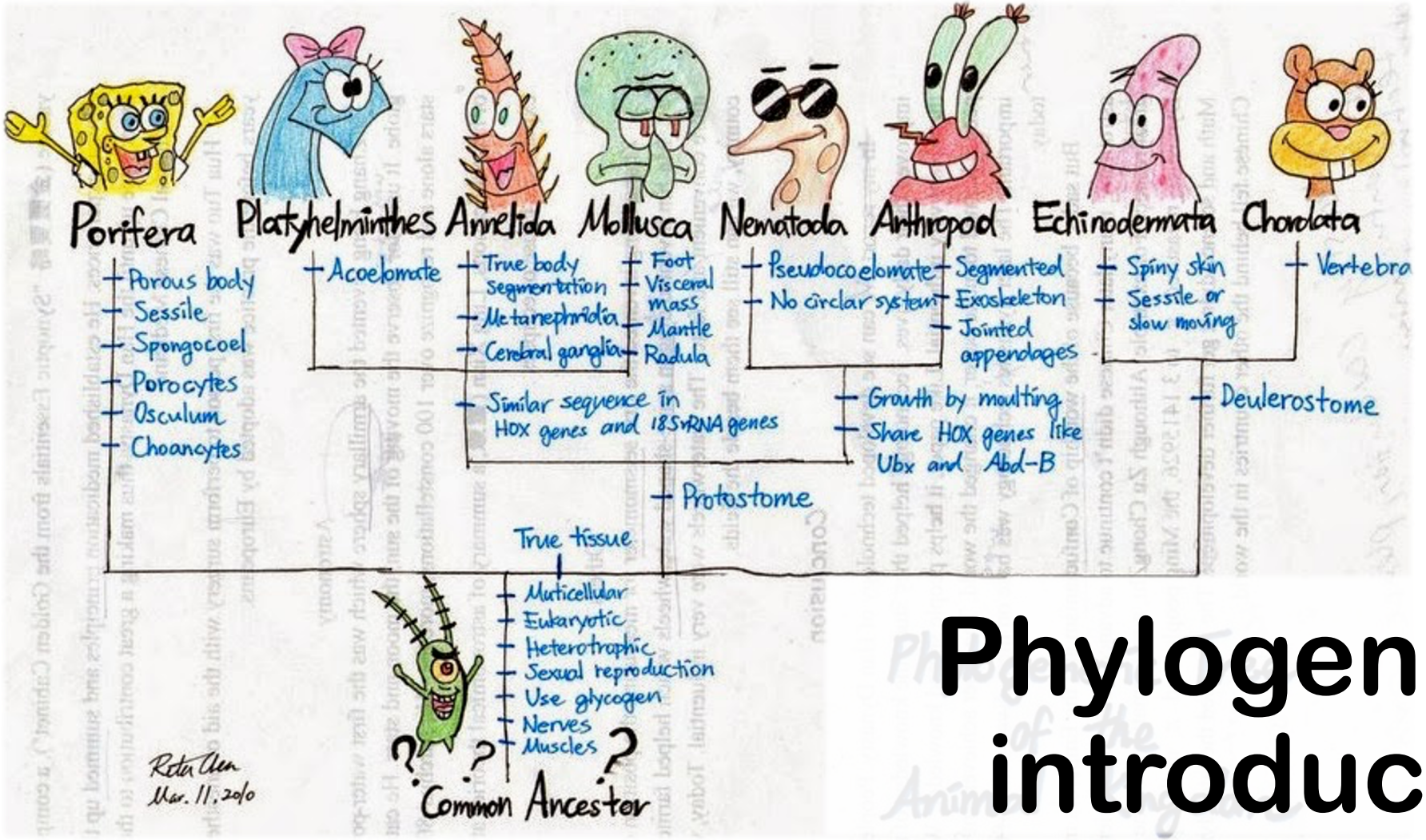


Phylogenetics introduction

http://phylonetworks.blogspot.com/2013/10/phylogenetics-with-spongebob.html
https://www.deviantart.com/thehurricanes/art/Phylogeny-with-Spongebob-156909686

Episode in the series on phylogenetics

Mikael Bodén

Part 1: Phylogenetics introduction

Motivation

Background

Homology

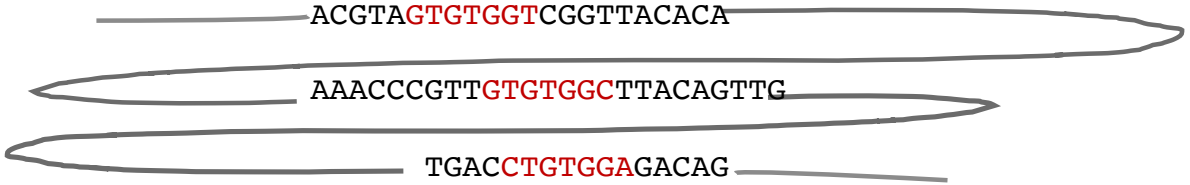
1

Reference …ACGTACGGTTACACAAAACCCGTTTTACGTAGTTGTGACG…

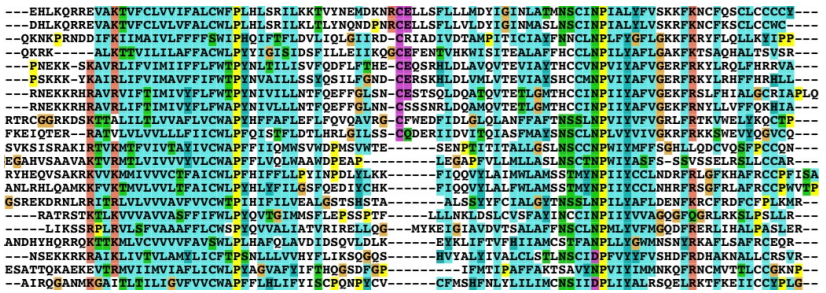
Sequence reads

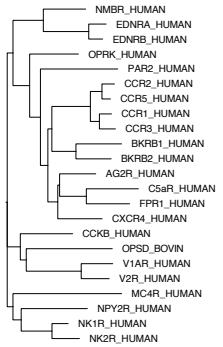
TTACACAAATCCCGTTGCACGTA
TACACAAATCCCGTTGTACGTAG
ACACAAAACCCGTTGCACGTAGT
CAAATCCCGTTGCACGTAGTTGT
AAATCCCGTTGTACGTAGTTGTG

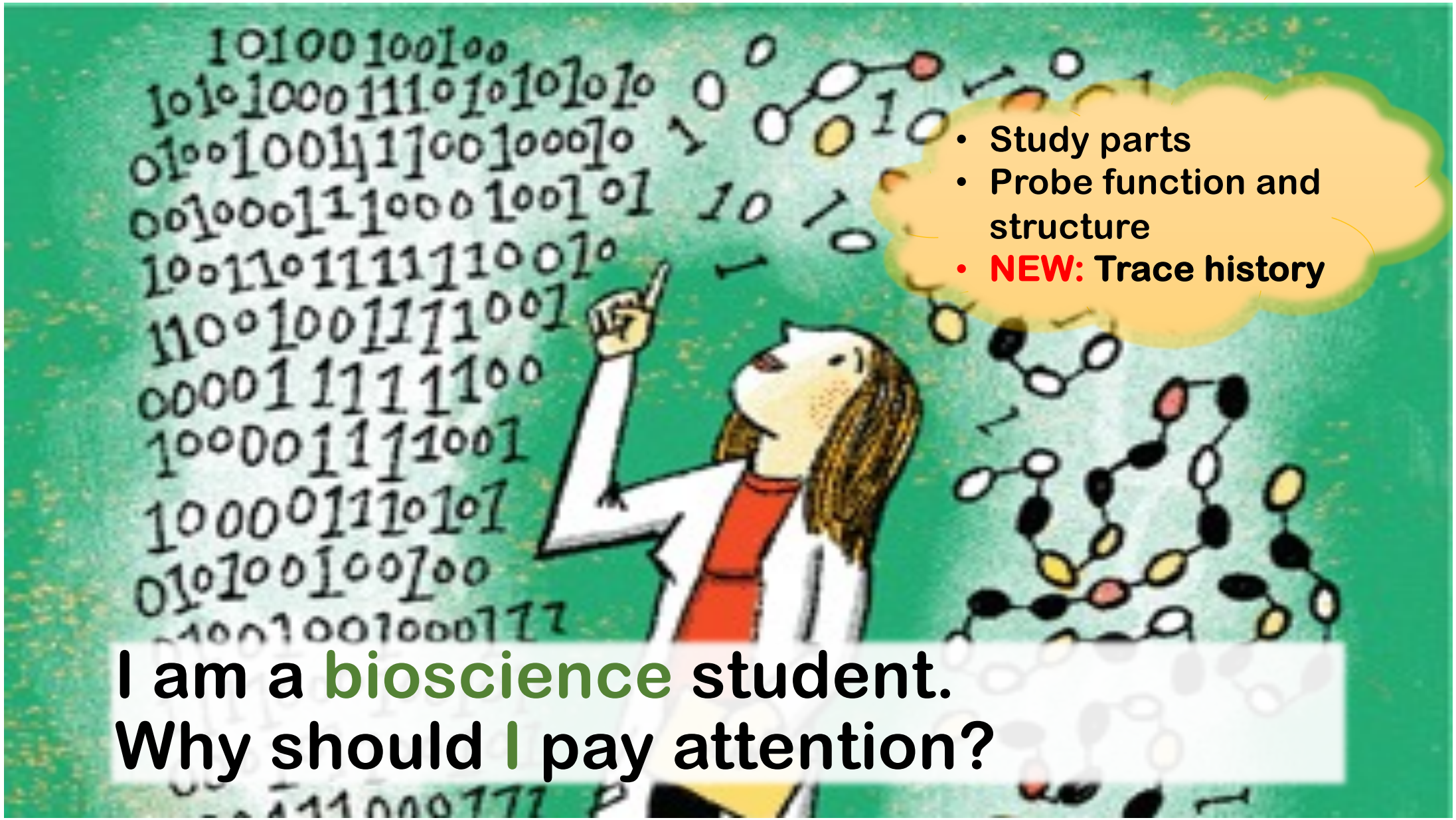
2



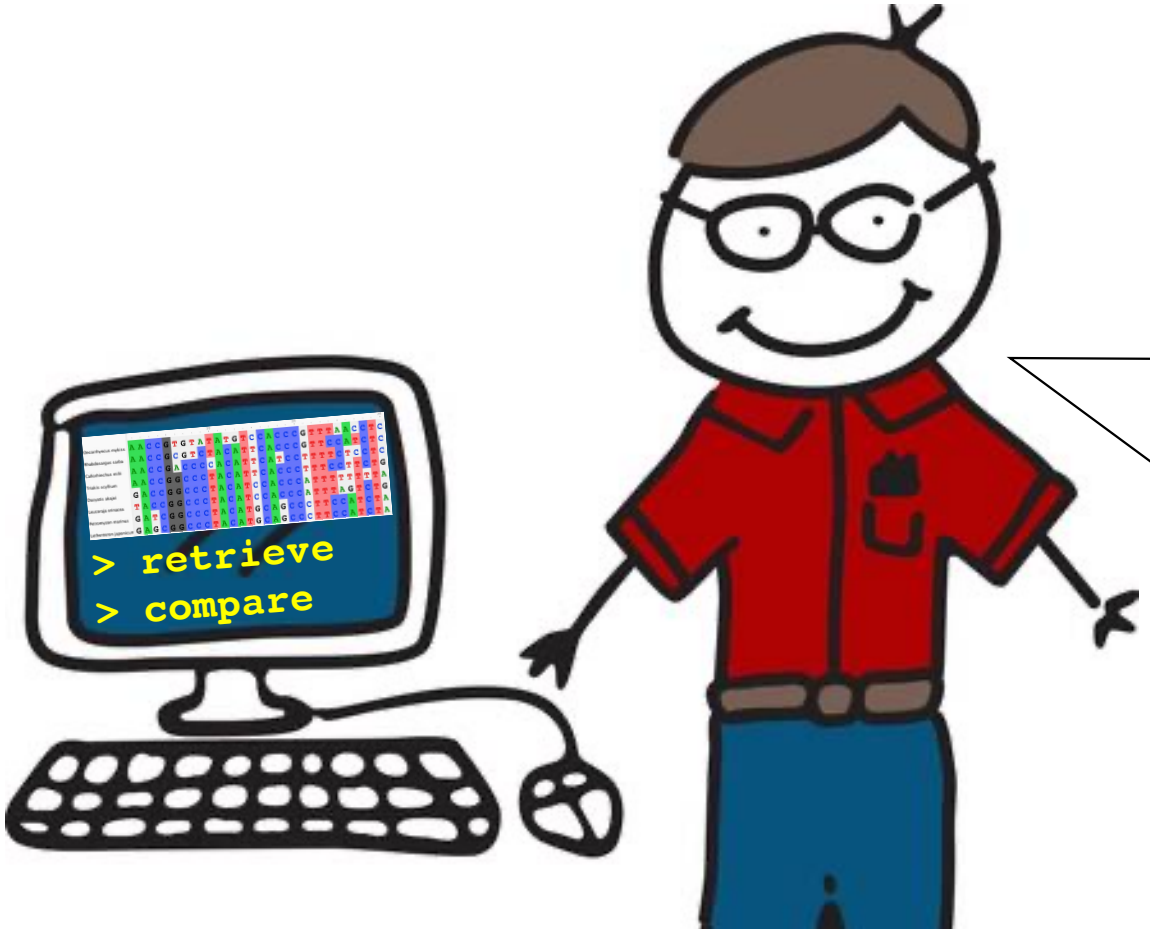
3






• Study parts
• Probe function and structure
• NEW: Trace history
I am a bioscience student.
Why should I pay attention?

I am a computing student. Why should I pay attention?

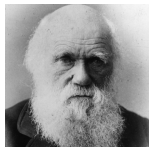


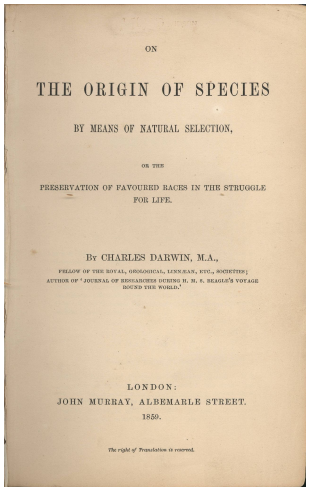
How cool is it that I will learn

- how to determine *when* existing sequences shared an ancestor?
- how to recreate phylogenetic trees, and ancient, probably extinct sequences?
- what it takes to identify the evolutionary events that formed existing sequences?

Phylogeny and common ancestor

- *All* organisms are related by genealogical descent with *modification*, as on a *branching* tree





ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

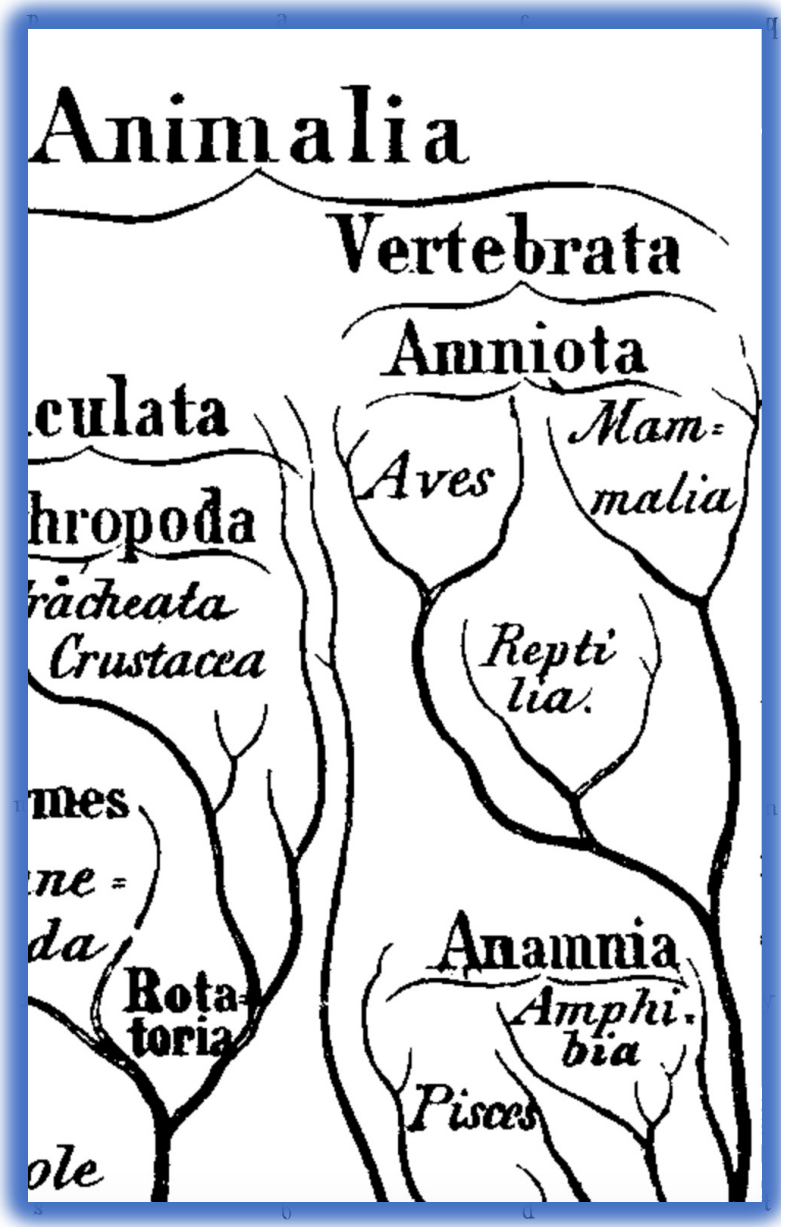
By CHARLES DARWIN, M.A.,

FELLOW OF THE ROYAL, GEOLOGICAL, LINNEAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

The right of Translation is reserved.

Phylogenetic tree suggested by Haeckel (1866) →



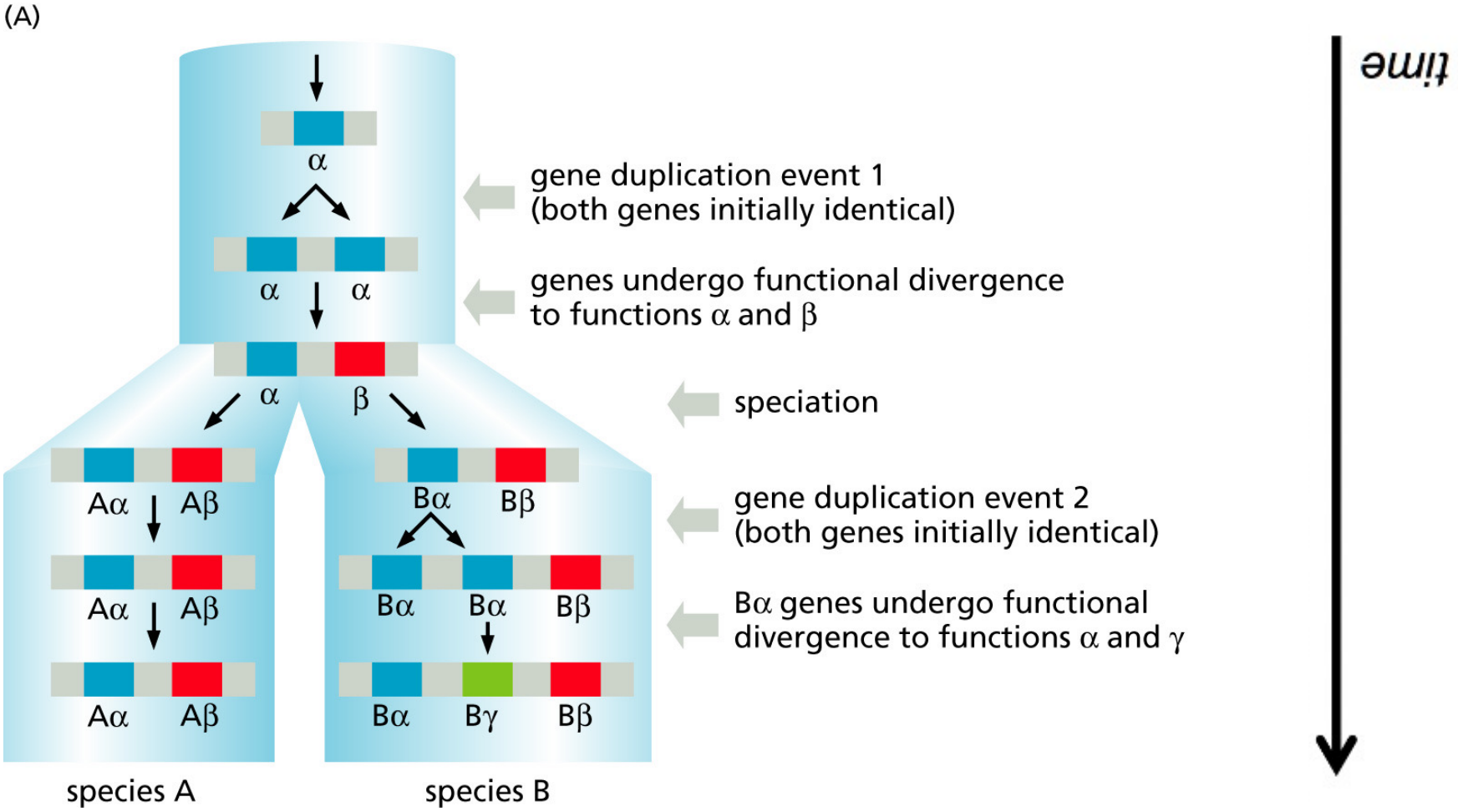
Homology, orthology and paralogy

Homology means descent from a common ancestor

Orthology is a special case of homology, in which the descent of sequences from a common ancestral form is topologically congruent with organismal speciation

Paralogy is a special case of homology, in which the sequences from a common ancestral sequence arose through duplication

Good reads:
Koonin EV (2001) *Genome Biol* 2(4); https://www.ncbi.nlm.nih.gov/pmc/articles/PMC138920/
Jensen RA (2001) *Genome Biol* 2(8); https://www.ncbi.nlm.nih.gov/pmc/articles/PMC138949/



Zvelebil & Baum, *Understanding Bioinformatics* Figure 7.10, p 242.

Part 2: Phylogenetics introduction

Sequence alignment

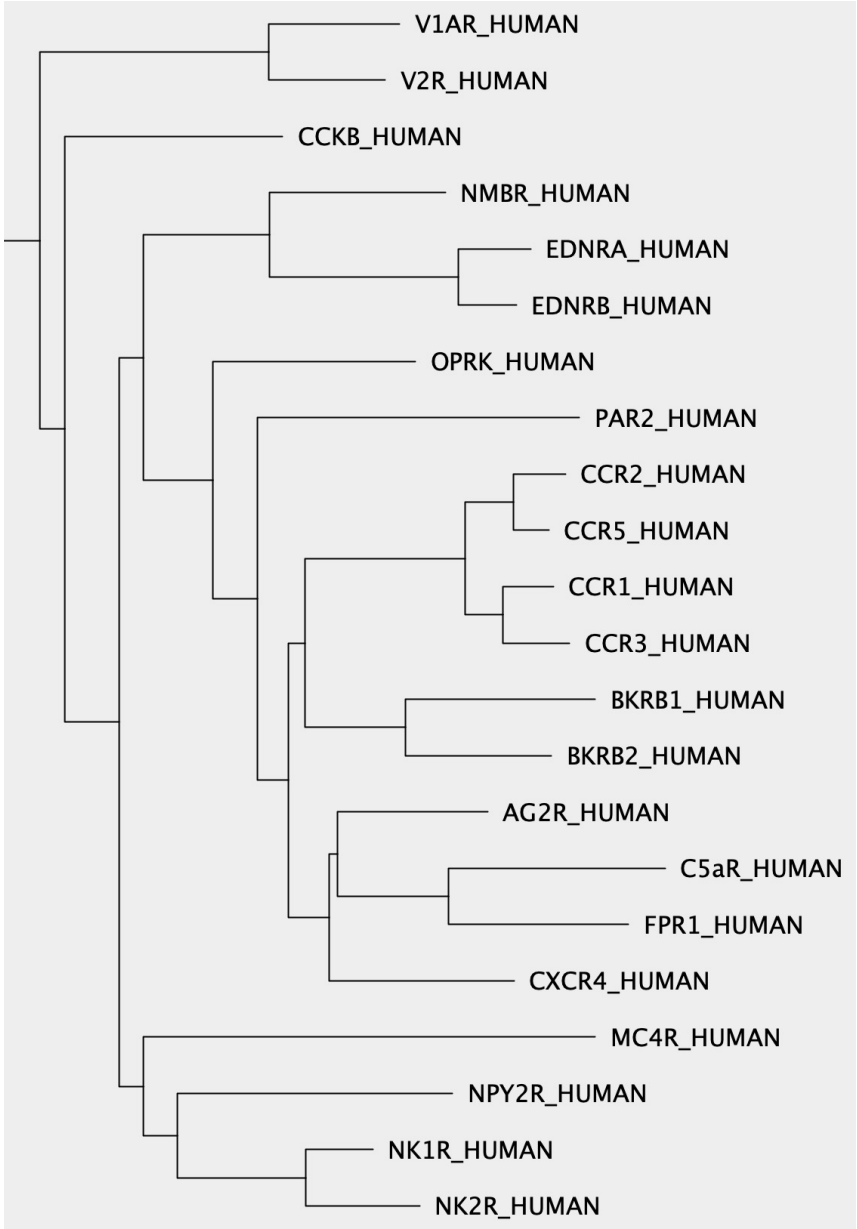
Phylogenetic trees, including terminology

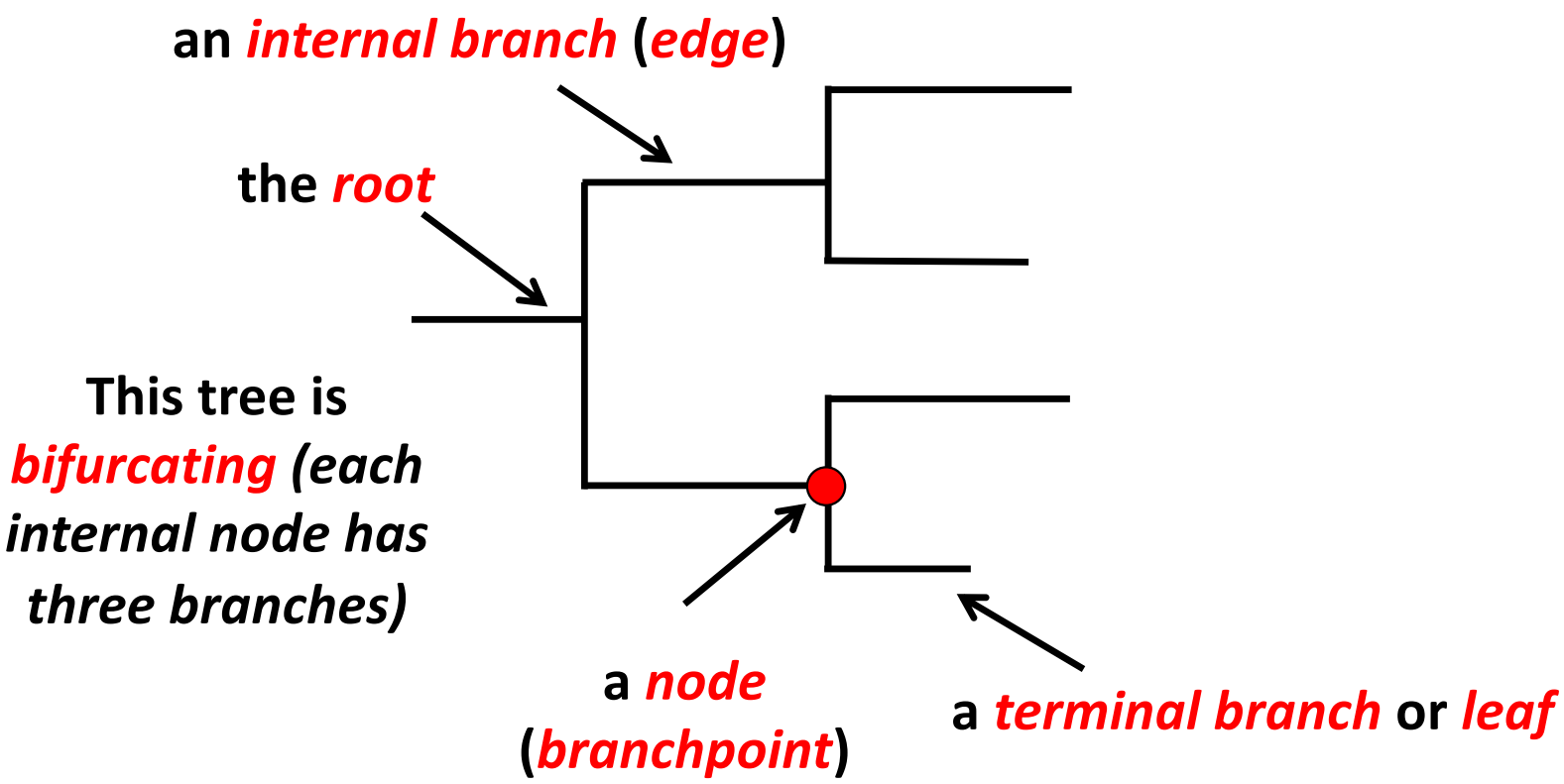
Species v. gene trees

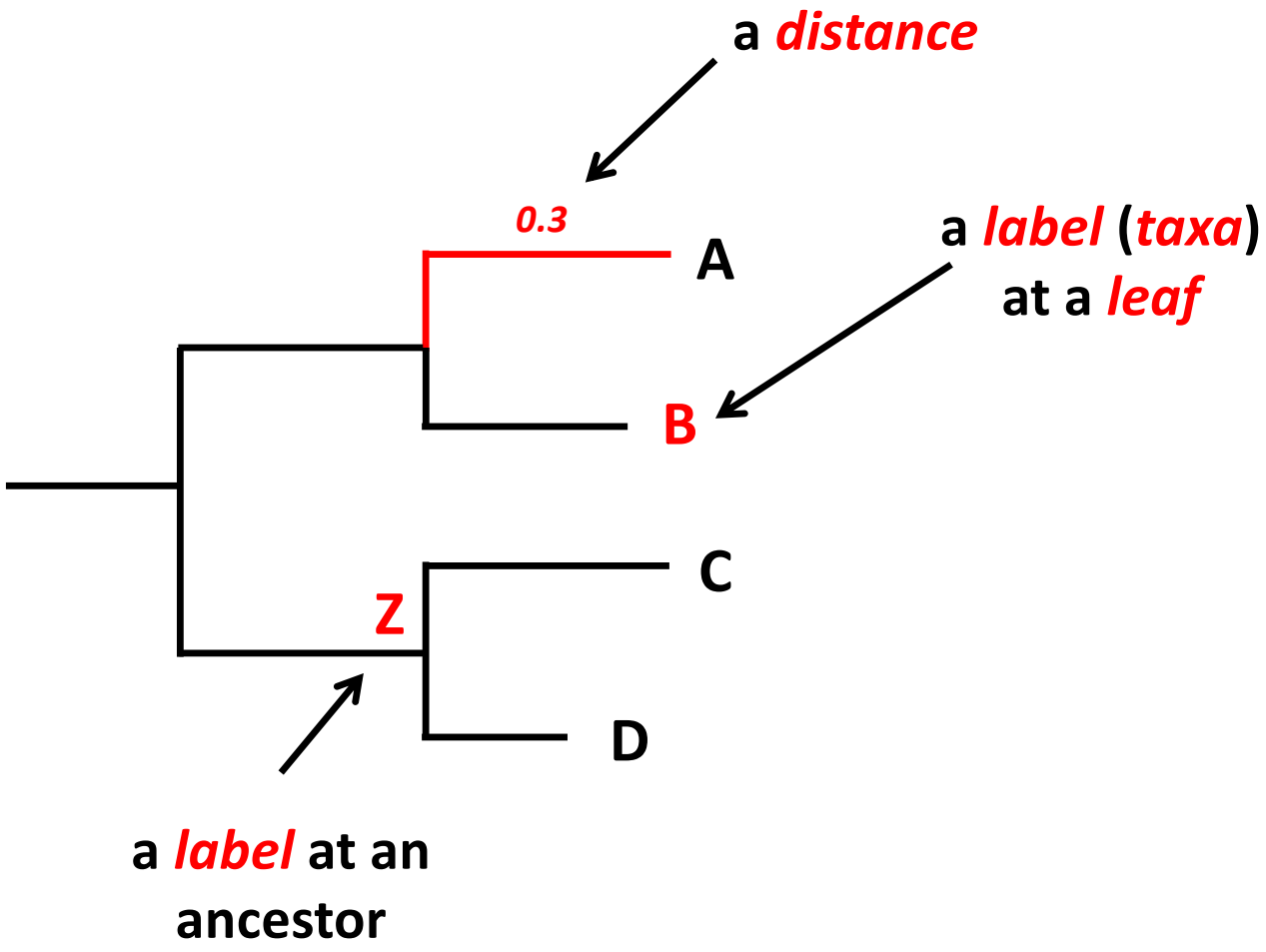
Multiple sequence alignment is input for phylogenetic analysis

Trees

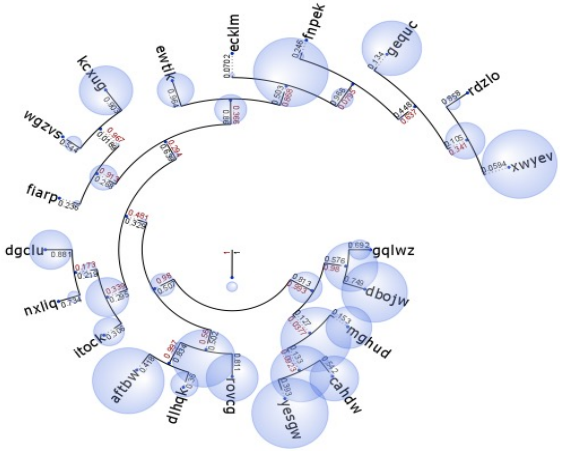
- Trees are connected graphs that do not contain circuits
- Trees are important as data structures
- A phylogenetic tree is a hypothesis about the evolutionary relationships between a set of objects (taxa); interpretation is subject to the type of data

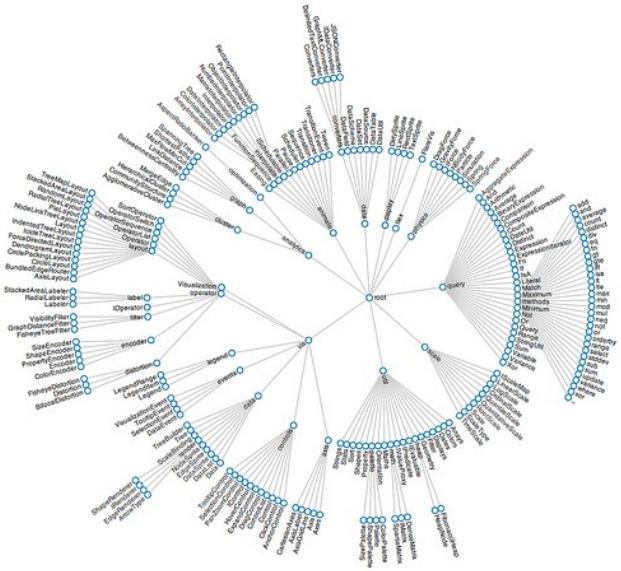


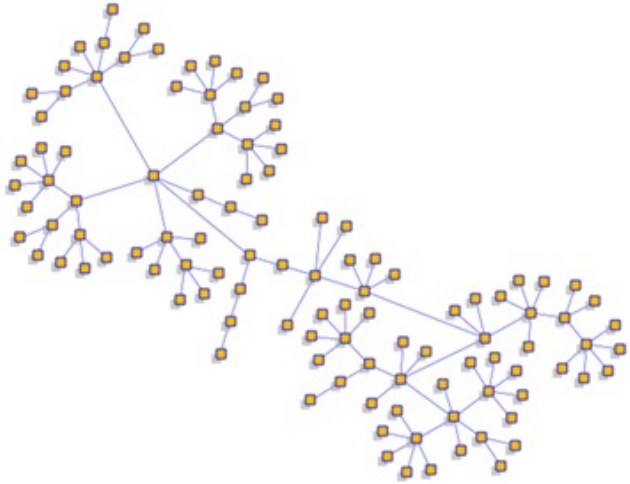

an internal branch (edge)
the root
This tree is bifurcating (each internal node has three branches)
a node (branchpoint)
a terminal branch or leaf

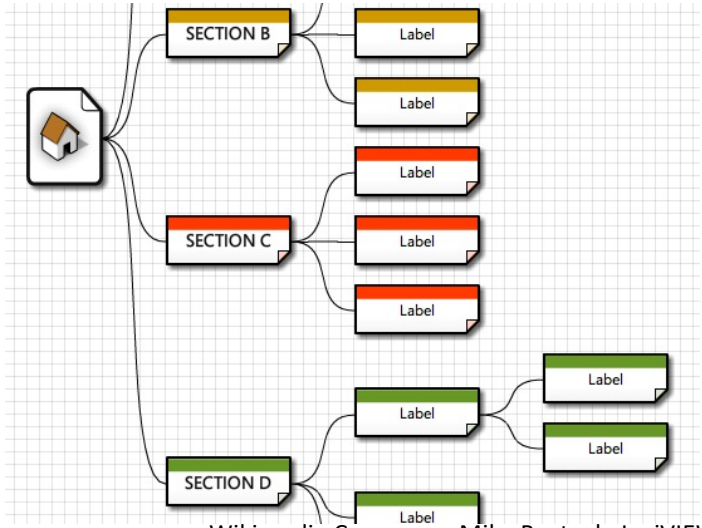


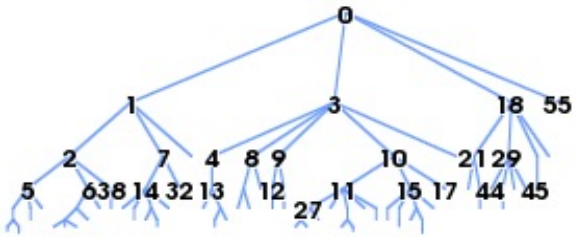
The vertical lines are NOT edges – they're purely for the viewer's convenience

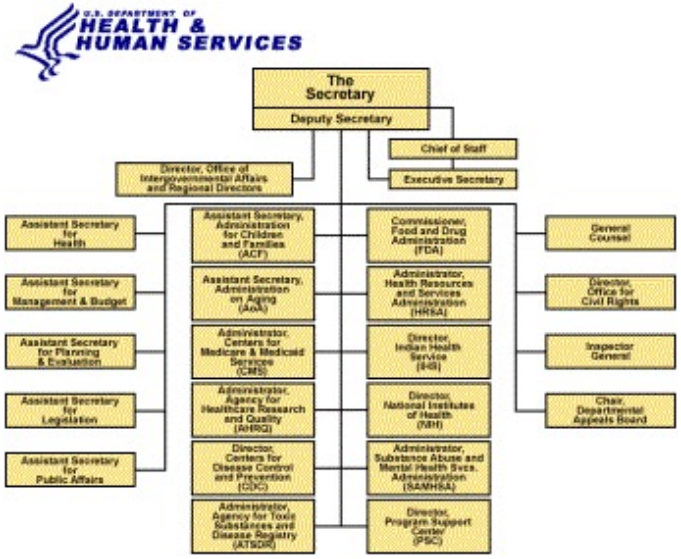








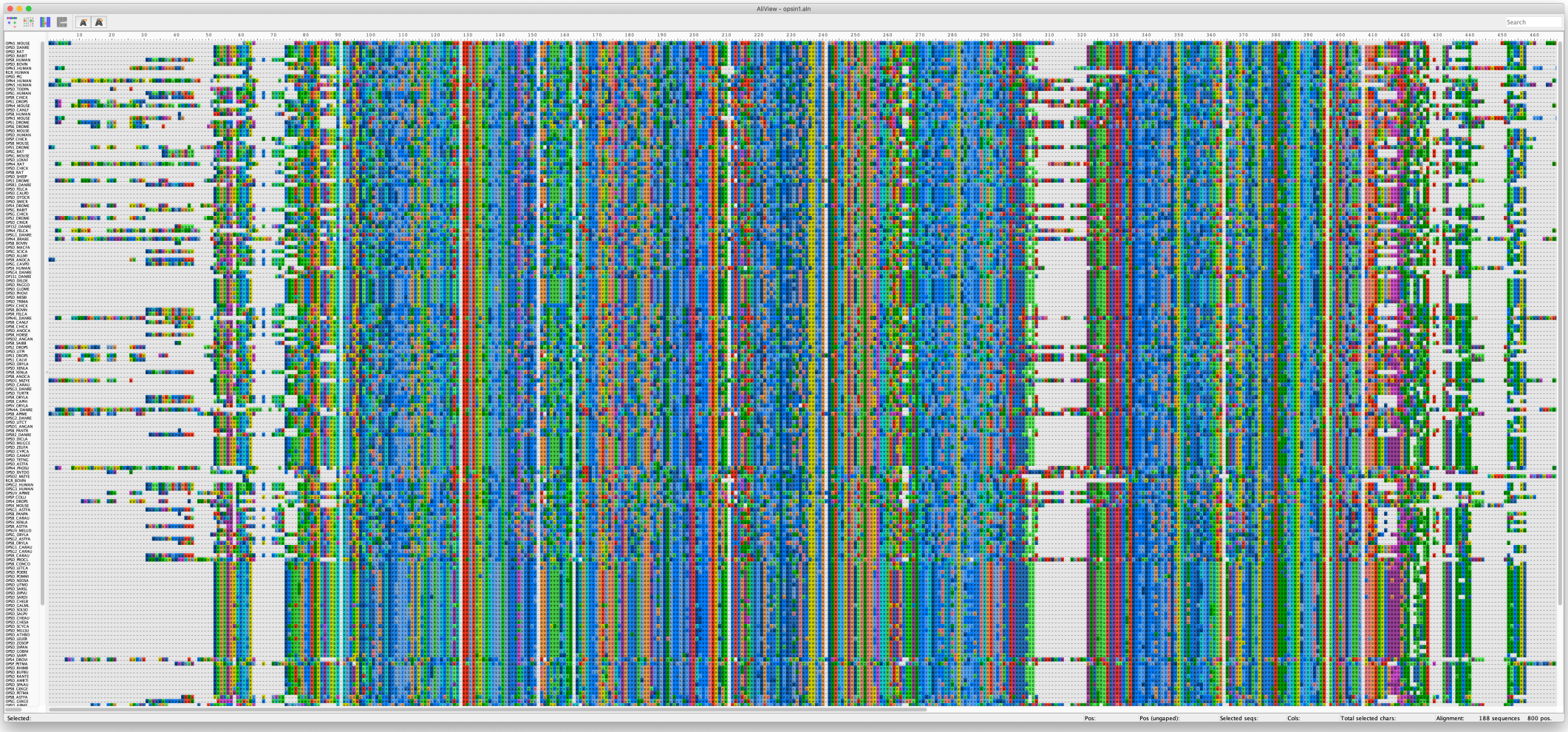


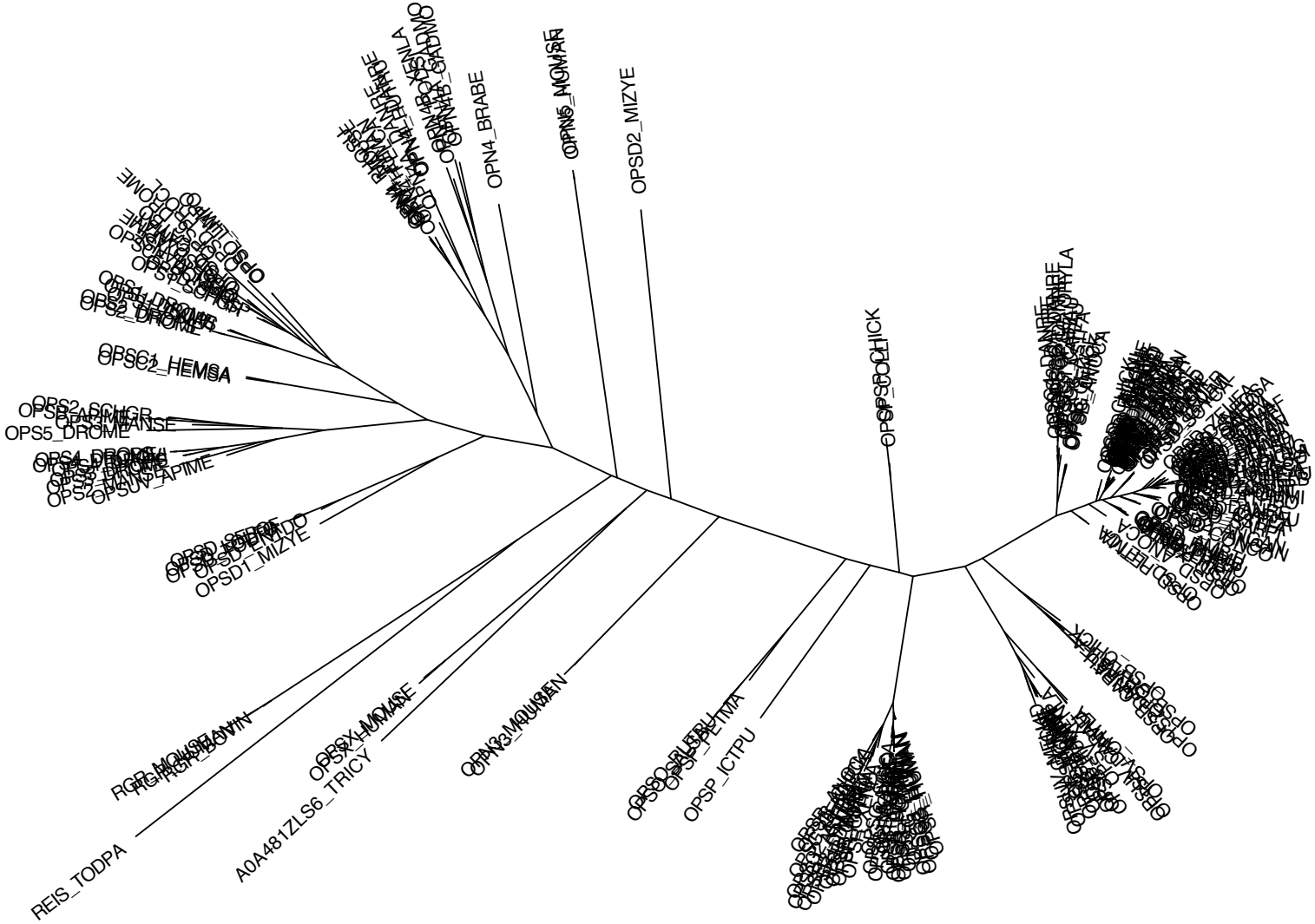


Wikimedia Commons; Mike Bostock; InviVIEW; LovelyCharts; Jaime Huerta-Cepas & Toni Gabaldón; US Dept Health & Human Services

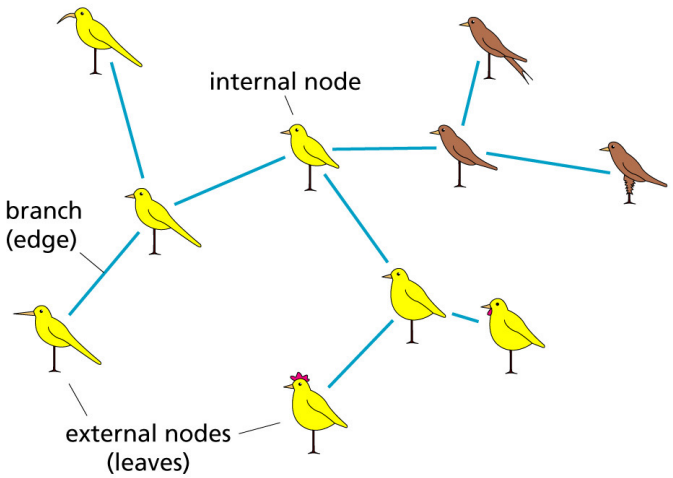
Opsin is a GPCR-1 sub-family

- Bilateria (16464 results)
  - Chordata (14310 results)
    - Branchiostoma (10 results)
      - Branchiostoma belcheri (Amphioxus) (5 results)
      - Branchiostoma floridae (Florida lancelet) (Amphioxus) (4 results)
      - Branchiostoma lanceolatum (Common lancelet) (Amphioxus lanceolatum) (1 results)
    - Vertebrata (14300 results)
      - Cyclostomata (jawless vertebrates) (30 results)
      - Gnathostomata (jawed vertebrates) (14270 results)
  - Protostomia (2154 results)
- Cnidaria (4 results)
  - Pocilloporidae (3 results)
  - Tripedalia cystophora (Jellyfish) (1 results)


AliView - opsin1.aln
Search
Selected:
Pos:
Pos (ungaped):
Selected seqs:
Cols:
Total selected chars:
Alignment:
188 sequences
800 pos.

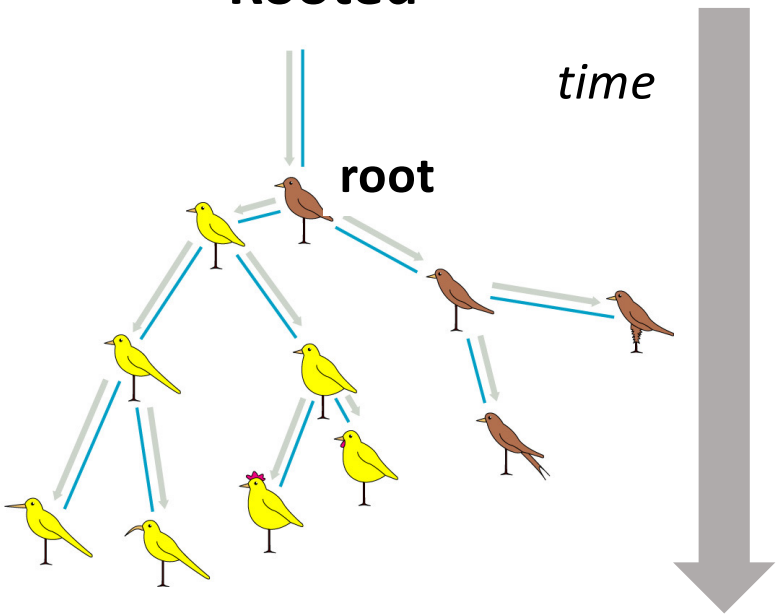

OPN4_BRABE
OPSD2_MIZYE
OPSD_CHICK
OPSP_ICTPU
OPSD_PETMA
OPSX_MOUSE
OPN3_MOUSE
A0A481ZLS6_TRICY
RGR_MOUSE
REIS_TODPA
OPSD_HEMSA
OPSU_APIME
OPS5_DROME

Unrooted



Distances are undirected

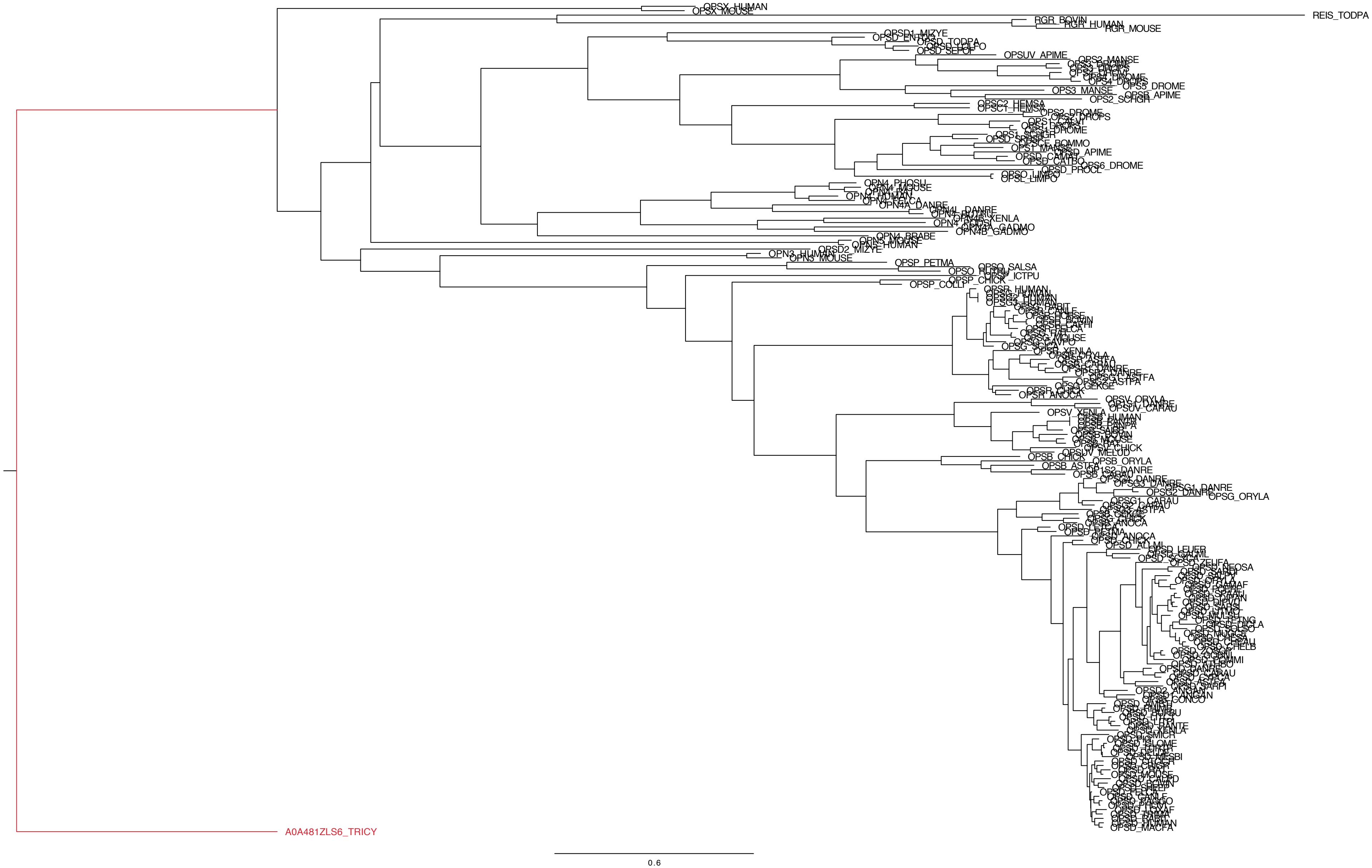
Rooted



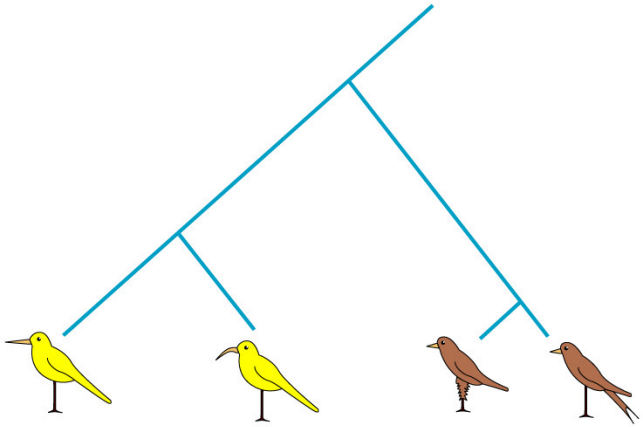
Distances are directed

Zvelebil & Baum, *Understanding Bioinformatics* p 227.

OPSD_FELCA
OPSD_LEUER
OPSD_PETMA
OPSB_CHICK
OPSB_ORYLA
OPSB_CARAU
OPSV_ORYLA
OPSUV_CARAU
OPSR_HUMAN
OPSR_ANOCA
OPSP_ICTPU
OPSO_SALSA
OPSO_RUTRU
OPN3_HUMAN
OPN3_MOUSE
A0A481ZLS6_TRICY
OPSX_HUMAN
OPSX_MOUSE
REIS_TODPA
RGR_BOVIN
RGR_HUMAN
RGR_MOUSE
OPSD1_MIZYE
OPSD_TODPA
OPSUV_APIME
OPS2_MANSE
OPS3_MANSE
OPS5_DROME
OPSB_APIME
OPS2_SCHGR
OPSC2_HEMSA
OPSC1_HEMSA
OPS2_DROME
OPS1_DROME
OPS6_DROME
OPSL_LIMPO
OPN4_RHOSU
OPN4_MOUSE
OPN4_HUMAN
OPN4A_DANRE
OPN4_XENLA
OPN4A_GADMO
OPN4B_GADMO
OPN4_BRABE
OPN5_MOUSE
OPN5_HUMAN
OPSD2_MIZYE
OPSP_CHICK
OPSP_COLLI
OPSG4_DANRE
OPSG_ORYLA
OPSG1_CARAU
OPSB_ANOCA
OPSD_CHICK
OPSD_ALLMI
OPSD_MACFA
0.6

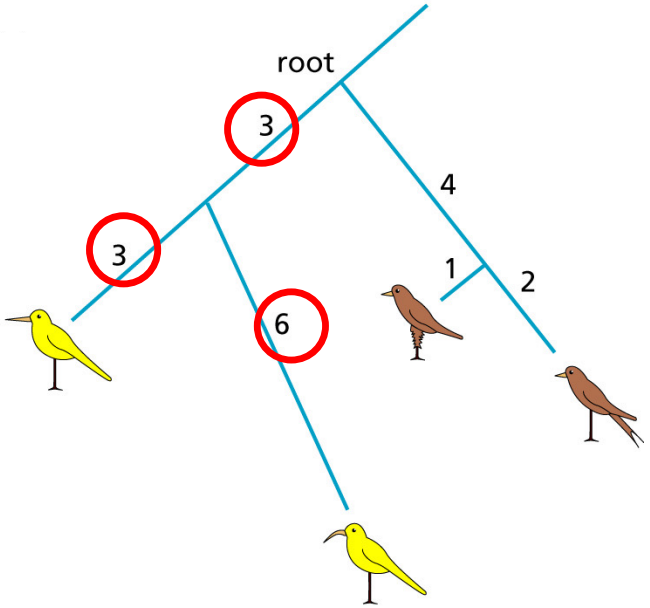


OPSX_HUMAN
OPSX_MOUSE
REIS_TODPA
RGR_BOVIN
RGR_HUMAN
RGR_MOUSE
OPSUV_APIME
OPS2_MANSE
OPS5_DROME
OPS3_MANSE
OPSB_APIME
OPS2_SCHGR
OPSC2_HEMSA
OPS6_DROME
OPN4_PHOSU
OPN4A_DANRE
OPN4B_GADMO
OPN3_MOUSE
OPSP_PETMA
OPSO_SALSA
OPSP_CHICK
OPSP_COLLI
OPSR_HUMAN
OPSR_ANOCA
OPSV_ORYLA
OPSUV_CARAU
OPSV_XENLA
OPSB_HUMAN
OPSB_CHICK
OPSB_ORYLA
OPSG1_DANRE
OPSG_ORYLA
OPSG1_CARAU
OPSD_PETMA
OPSD_ANOCA
OPSD_ALLMI
OPSD_LEUER
OPSD2_ANOAN
OPSD_MACFA
A0A481ZLS6_TRICY
0.6



Cladogram

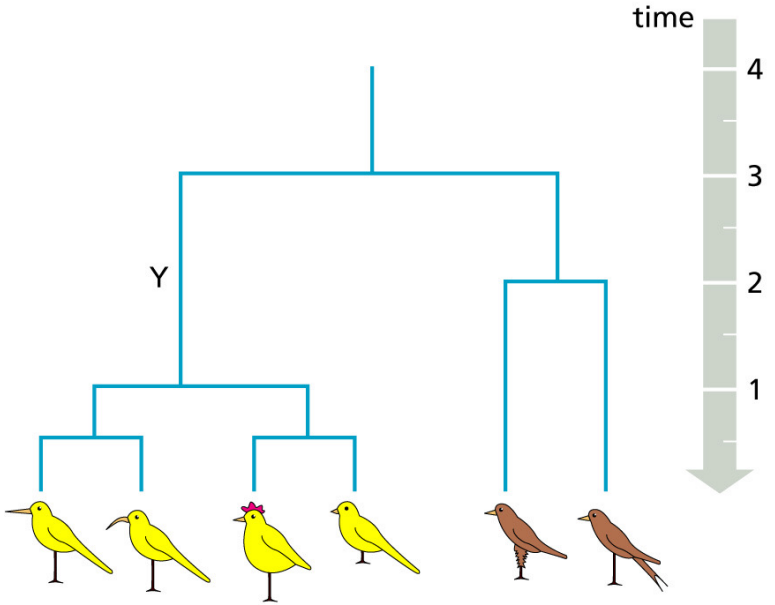
branch lengths have *qualitative*
but not *quantitative* meaning



Additive tree

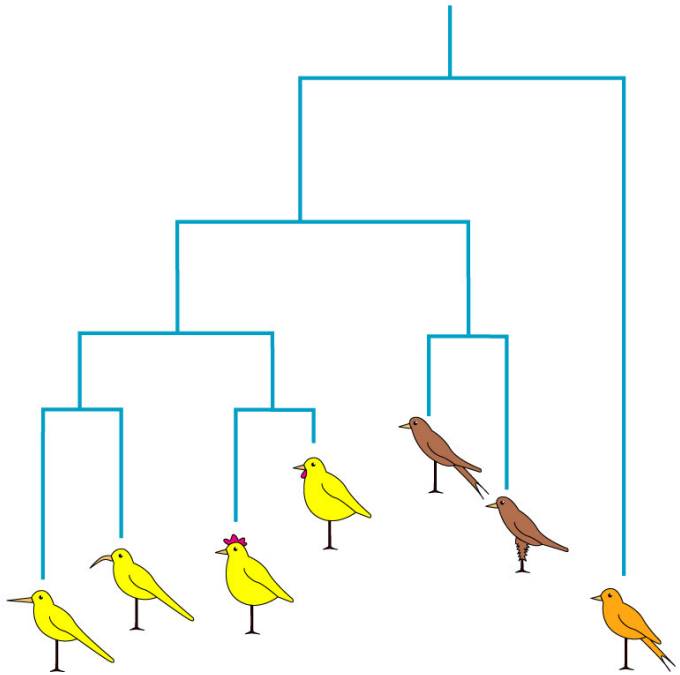
branches quantify the evolutionary
distance, the progress of divergence

Zvelebil & Baum, *Understanding Bioinformatics* p 229.



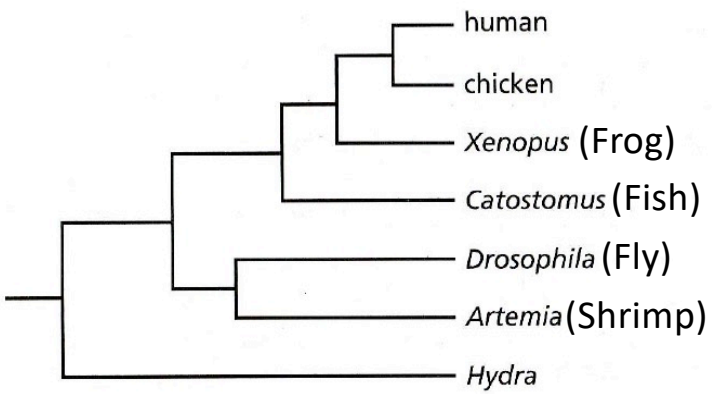
Ultrametric tree

(constant rate of mutation; *molecular clock*)



Additive tree with outgroup

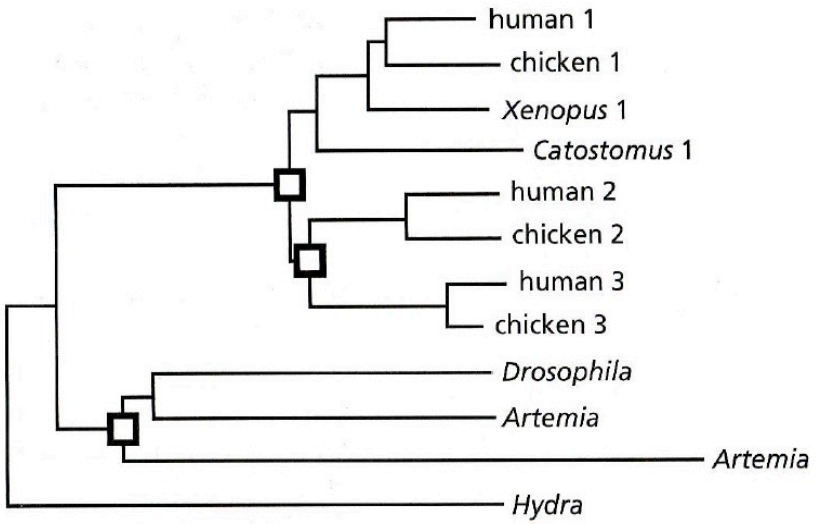
Zvelebil & Baum, *Understanding Bioinformatics* p 229.



Species tree

(evolution of species)

Ideally derived from **strictly orthologous** sequences

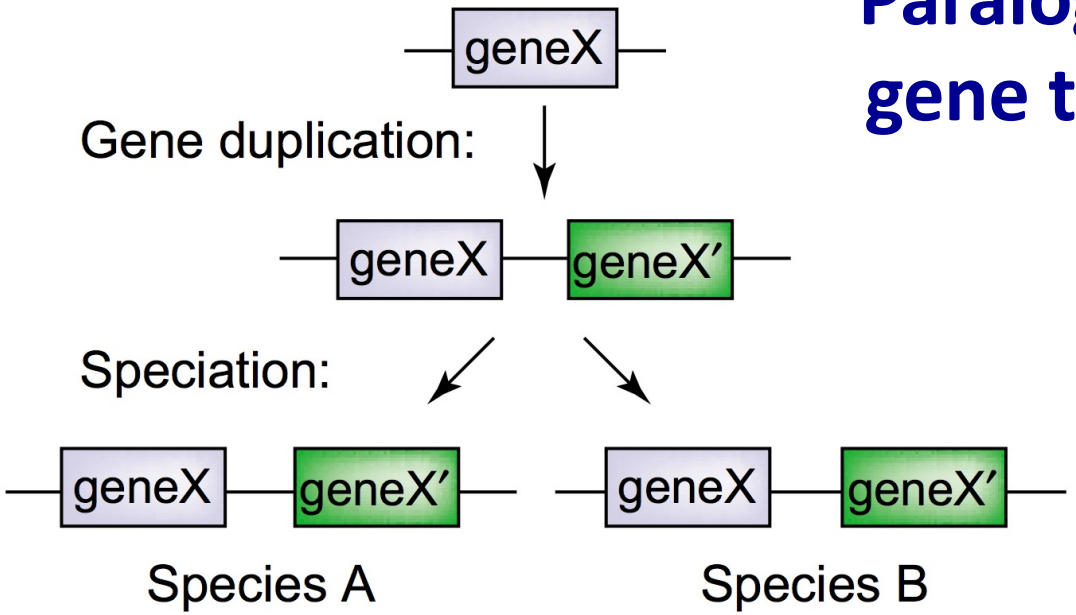


Gene tree

(evolution of homologous genes, in this case a particular family of membrane proteins)

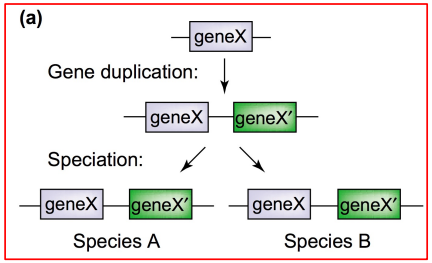
Zvelebil & Baum, *Understanding Bioinformatics* p 231.

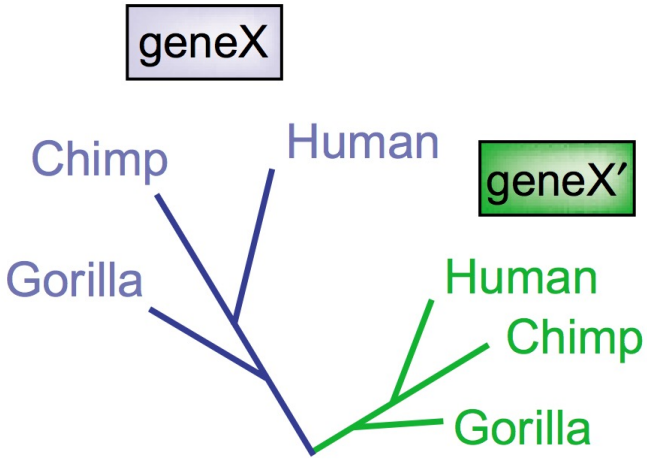
Paralogs in gene trees



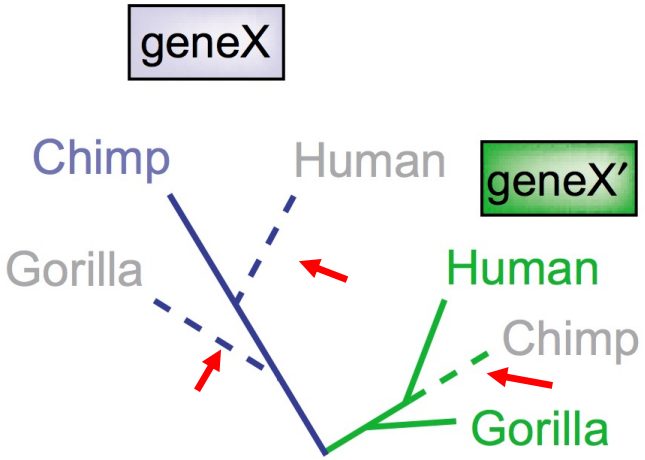
Gene X is duplicated prior to speciation. All subsequent species inherit both copies of the gene (unless one or the other is lost somewhere along the way).

Baldauf, SL (2003) *Trends in Genetics* 19: 345-351.





All sequences of gene X are **orthologs** of each other, and all the sequences of gene X' are **orthologs** of each other. However, X and X' are **paralogs**. Both the X and X' subtrees show the true relationships among the three species.



A tree of the X/X' gene family can be misleading if not all the sequences are included (because of **incomplete sampling** or **gene loss**). If the broken branches are missing, then the true species relationships are misrepresented.

Read more about this?

- Some of the material is based on Zvelebil and Baum's textbook "Understanding Bioinformatics" (chapters 7 and 8)