ISC 5317 / ISC 4933

CEB COMPUTATIONAL EVOLUTIONARY BIOLOGY

Class Meeting

Lectures:
Tuesdays and Thursdays 11:00AM-12:15 PM Dirac Science Library Room 152

Instructor

Peter Beerli
Office: 150-T DSL
Email: beerli@fsu.edu
Phone: (850) 559-9664

Class Assistant

Kyle Shaw
Office: 150-J DSL
Email: shawk3@outlook.com
Phone: TBA

Office Hours

- Peter Beerli: by appointment (email: beerli@fsu.edu or text to 850 559 9664); or just come to my office, If do not have a meeting I will have time for you.
- Kyle Shaw: we will set up an open lab session every week so that students can get help on python and the assignments.

Objectives

This course will introduce students to methods used in phylogenetics and population genetics and writing computer programs using such methods. Primary objectives of the course are:

1. to expose students to a large set of modern methods used in the field of theoretical evolutionary biology, and learn about the details of often used methods in phylogenetic analysis and population genetics analysis.

2. to introduce students to the programming aspects of the field. Students will learn and use the programming language Python to develop scripts and to understand details of the methods.

3. to empower students to develop programming and analysis skills that involve development of scripts to change data format, execute applications, and analyze results.

Content

Advanced computational methods are becoming increasingly important in biology. A wide range of applications — including, for instance, identifying pathogens, tracing viral transmission pathways, and reconstructing the geographic expansion of humans out of Africa — rely on evolutionary inference. This course will cover the methods currently used for evolutionary inference, the stochastic models and inference principles they are based on, and how they are implemented in practice. The students will get hands-on experience in developing computational software implementing these methods. We expect that the students leave the course with the necessary skills to develop their own ideas and are able to develop projects that are based on simulated data sets and scripts.

Grading

- Grades will be based on students' execution of the 5 (programming) assignments, each of which involves understanding the algorithms, code design, and program documentation [100 points each]
- Either two students or a single student will work on a project on their own during the last 4 weeks of the semester and give a short presentation of their work during the last second two classes periods. I expect that group projects are twice as large as single student projects [100 points for the report and 100 points for the presentation]
- We will have a theory test on November ~~7th~~ 9th (midterm). [100 points]
- There will be no final exam, the project substitutes for a final examination. The total number of points is 800.

A student who accumulates 90% or more of the possible points will receive a grade of "A", a student who accumulates between 80% and 89% of the possible points will receive a grade of "B", a student who accumulates between 70% and 79% of the possible points will receive a grade of "C", a student who accumulates between 60% and 69% of the possible points will receive a grade of "D", and a student who accumulates less than 60% of the possible points will receive a grade of "F".

Missed/Late Assignments

Deadlines for assignments will be announced in class; late assignments will be accepted for full grade only in cases of illness or death in the family. 5% of the total points (100pt) are deducted per day for late assignments.

Lectures: Topic overview

1. Processes and patterns
    - Population genetics: Wright-Fisher population models, coalescence theory;
    - Phylogenetics: tree structures, speciation, Gene tree versus Species tree
    - Mutation models: mutation/substitution model
    - Simulation of data
2. Inference:
    - Parsimony and Distance methods
    - Maximum likelihood, Bayesian inference, Monte Carlo, Markov chain Monte Carlo,
    - Model selection
    - Bootstrap/Jacknife

Each topic will include computational algorithms, problematic aspects such as convergence, biases, main focus will be on Bayesian and maximum likelihood methods.

Assignments

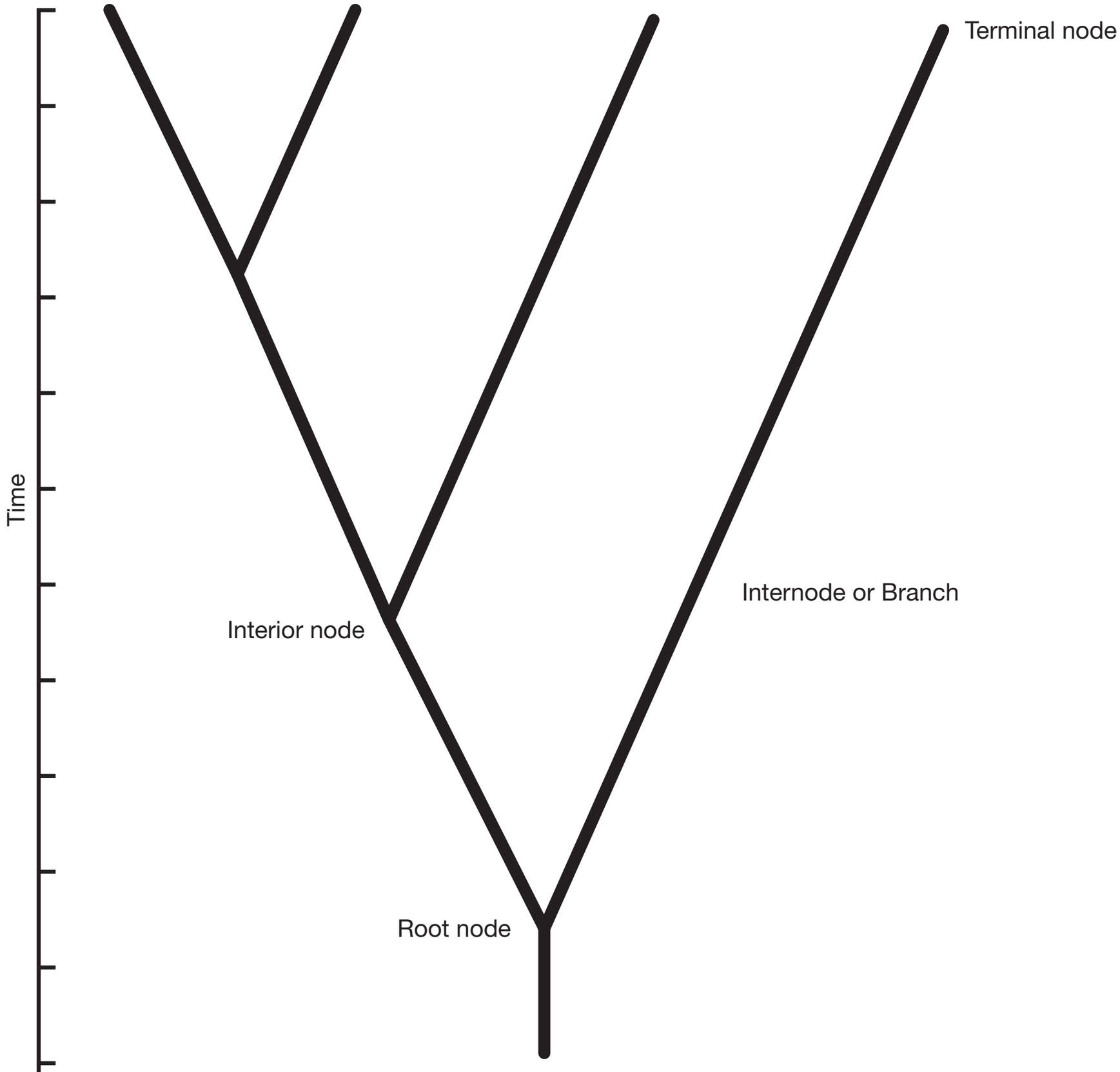
This list of assignments is an example, difficulty of assignments will depend on the overall class programming skills. Each assignment topic will be introduced in detail during class. The final set of assignments is not specified yet but it will look similar to the ones shown below:

1. Read and write a tree structure
2. Simulate data on a tree
3. Simulate data using the coalescent
4. Construct an ABC sampler to estimate the effective population size
5. Model selection using the program MIGRATE
6. Project: The project will discuss either (1) a complex analysis of data or (2) software development or (3) a theory section we did not discuss. The project consists of two parts, a report (of not more than 8 pages) and a presentation of 10 minutes. We will develop ideas for the project during class.

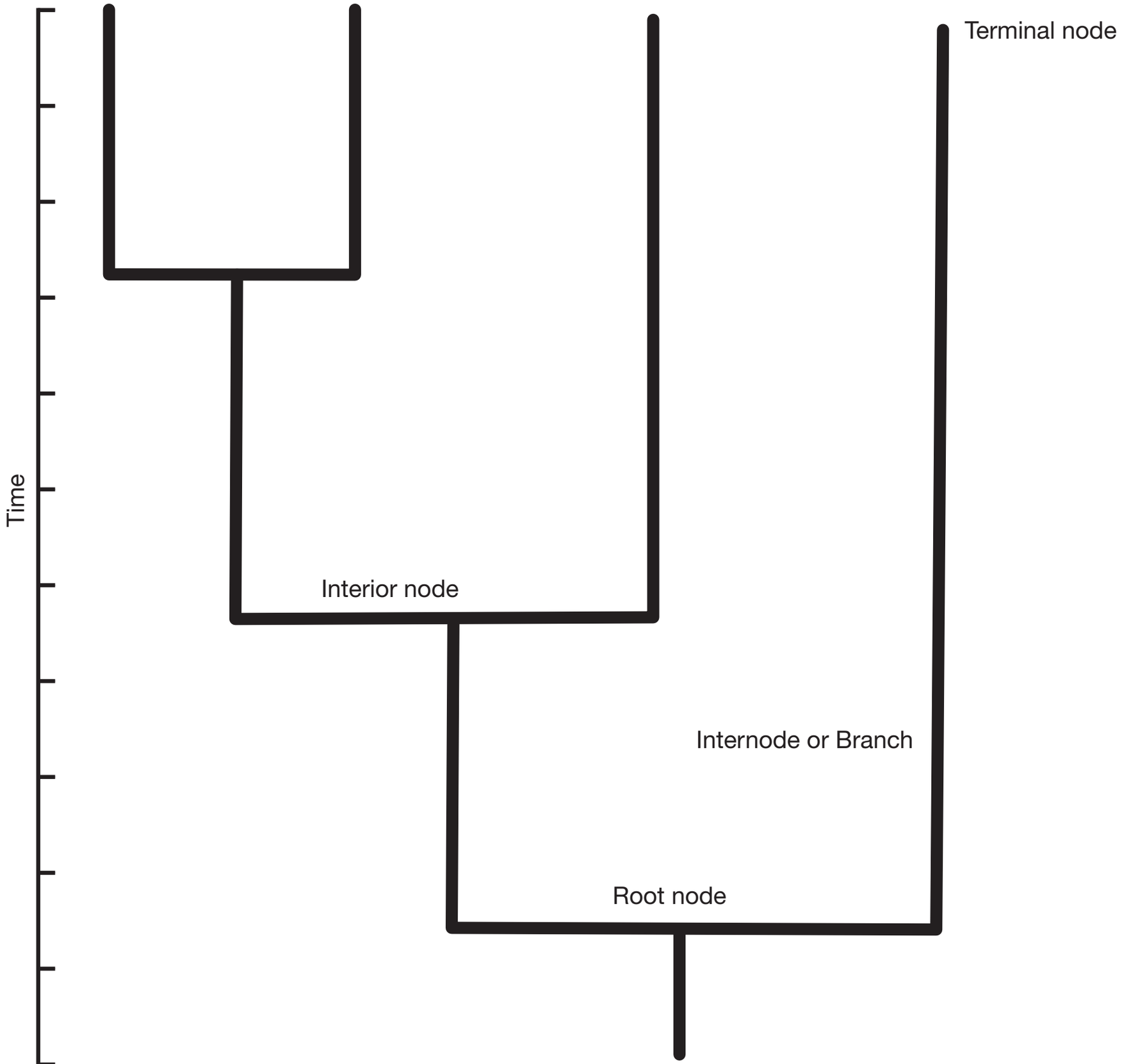
Lecture Schedule

1. Introduction. Trees and tree representation (Aug. 29)
2. Python and trees (Aug. 31)
3. Parsimony (Sep 5)
4. Python and parsimony (Sep 7)
5. Searching for the best tree(s) (Sep 12)
6. Substitution models and distance (Sep 14)
7. Substitution models general Sep 19)
8. Substitution model exercise Sep 21)
9. Rate variation and more substitution models Sep 26)
10. Maximum Likelihood (Sep 28)
11. Maximum Likelihood (Oct 3)
12. Bayesian inference (Oct 5)
13. Markov chain Monte Carlo (Oct 10)
14. ABC – approximate Bayesian computing (Oct 12)
15. The coalescent (Oct 17)
16. Coalescent simulation and extensions to the coalescent (Oct 19)
17. Gene tree vs Species tree (Oct 24)
18. Gene tree vs Species tree (Oct 26)
19. Model Selection (Oct 31)
20. Bootstrap/Jacknife (Nov 2)
21. Review session (Nov 7)
22. Mid term (Nov 9)
23. Project (Nov 14)
24. Project (Nov 16)
25. Project (Nov 21)
26. Project (Nov 28)
27. Project (Nov 30)
28. Presentation (Dec 5)
29. Presentation (Dec 7)

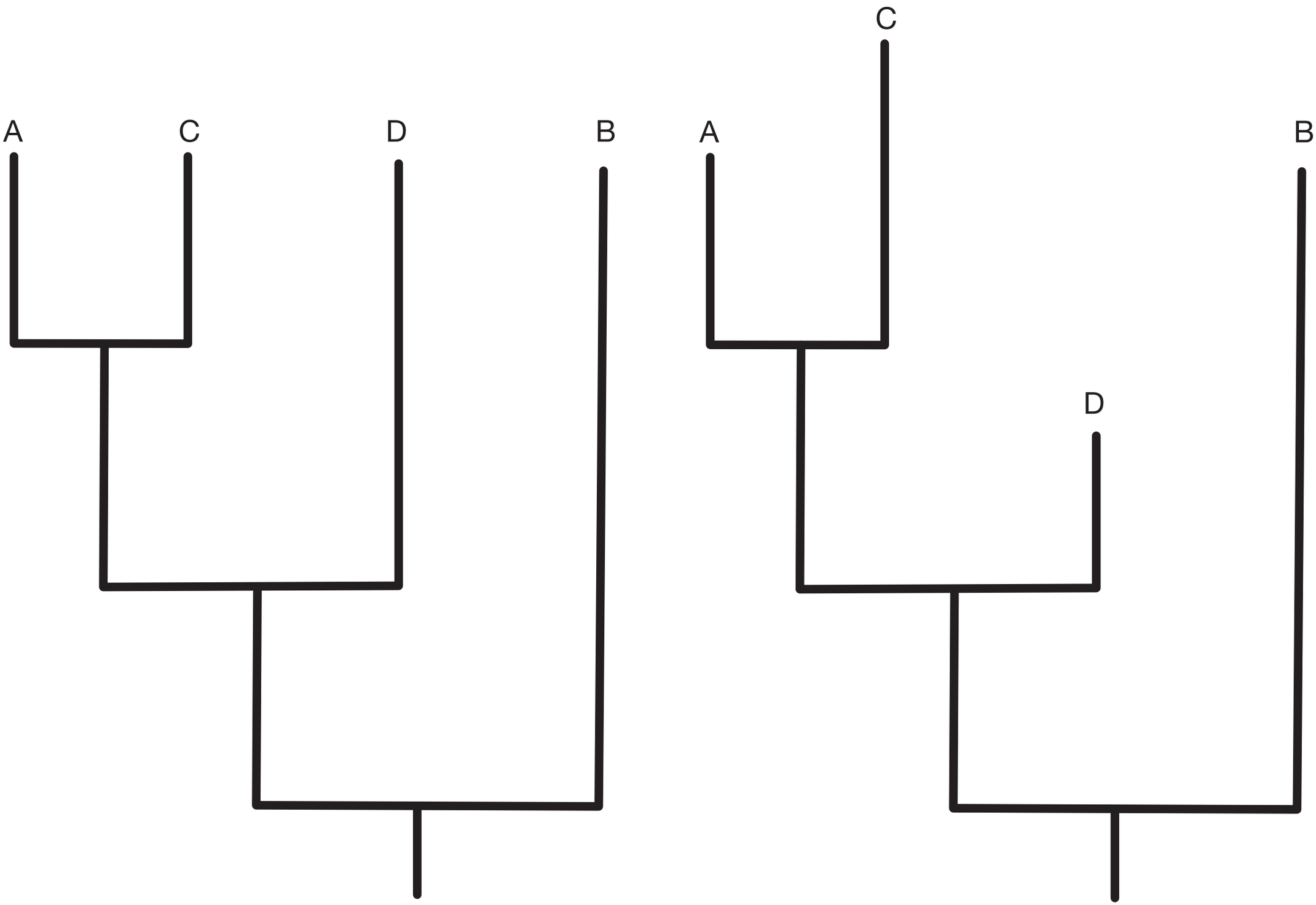
PHYLOGENETIC TREES



PHYLOGENETIC TREES

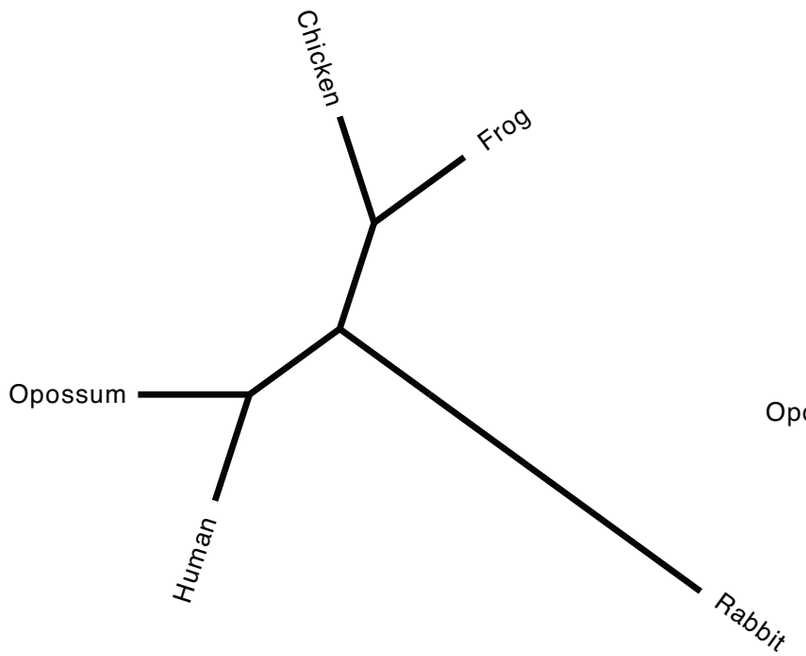


PHYLOGENETIC TREES

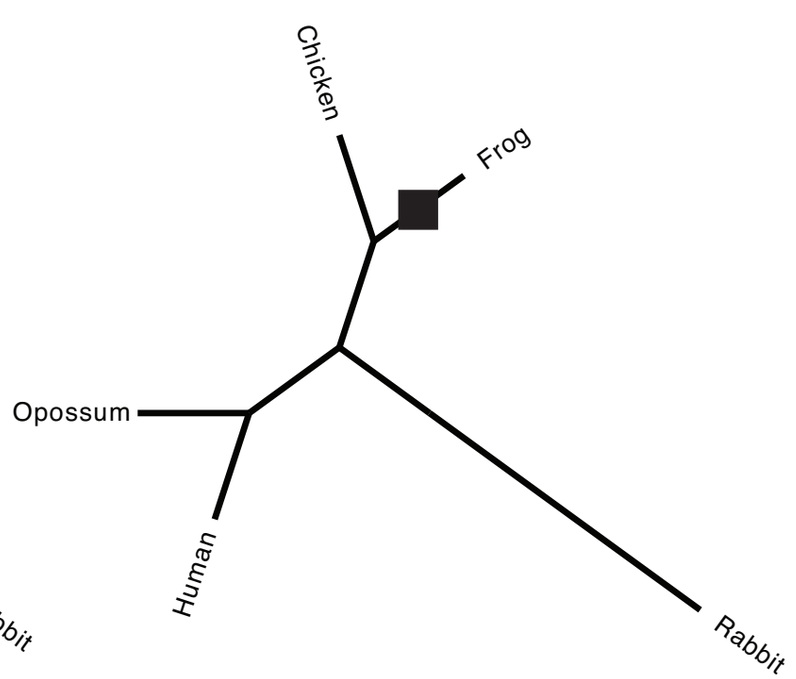


PHYLOGENETIC TREES

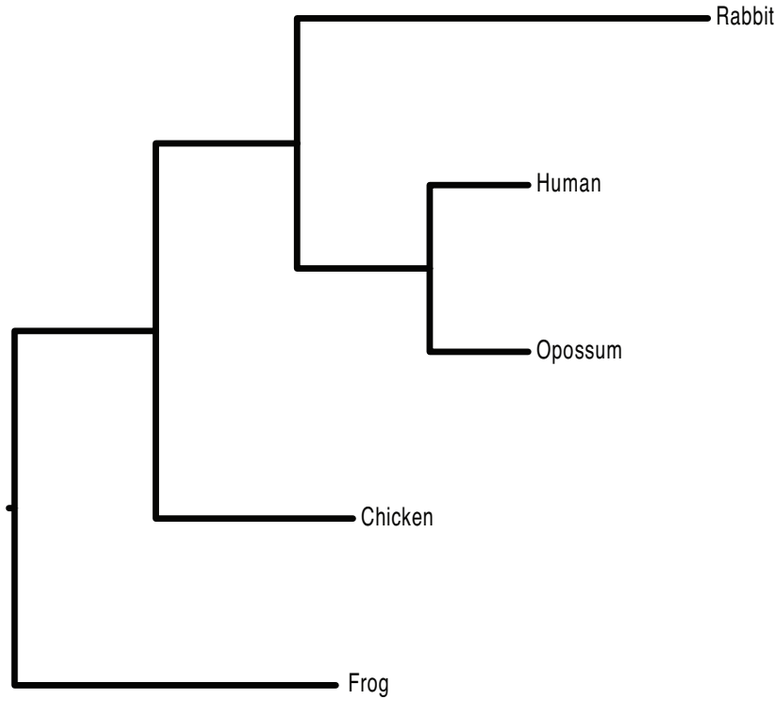
a



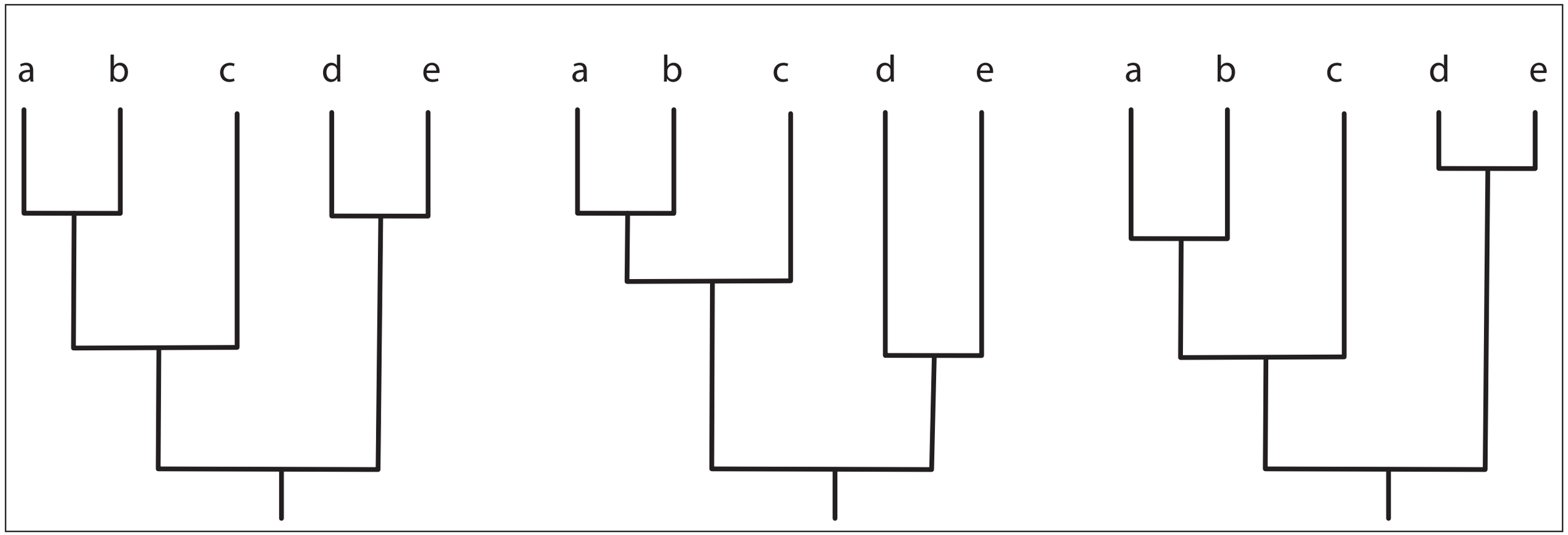
b

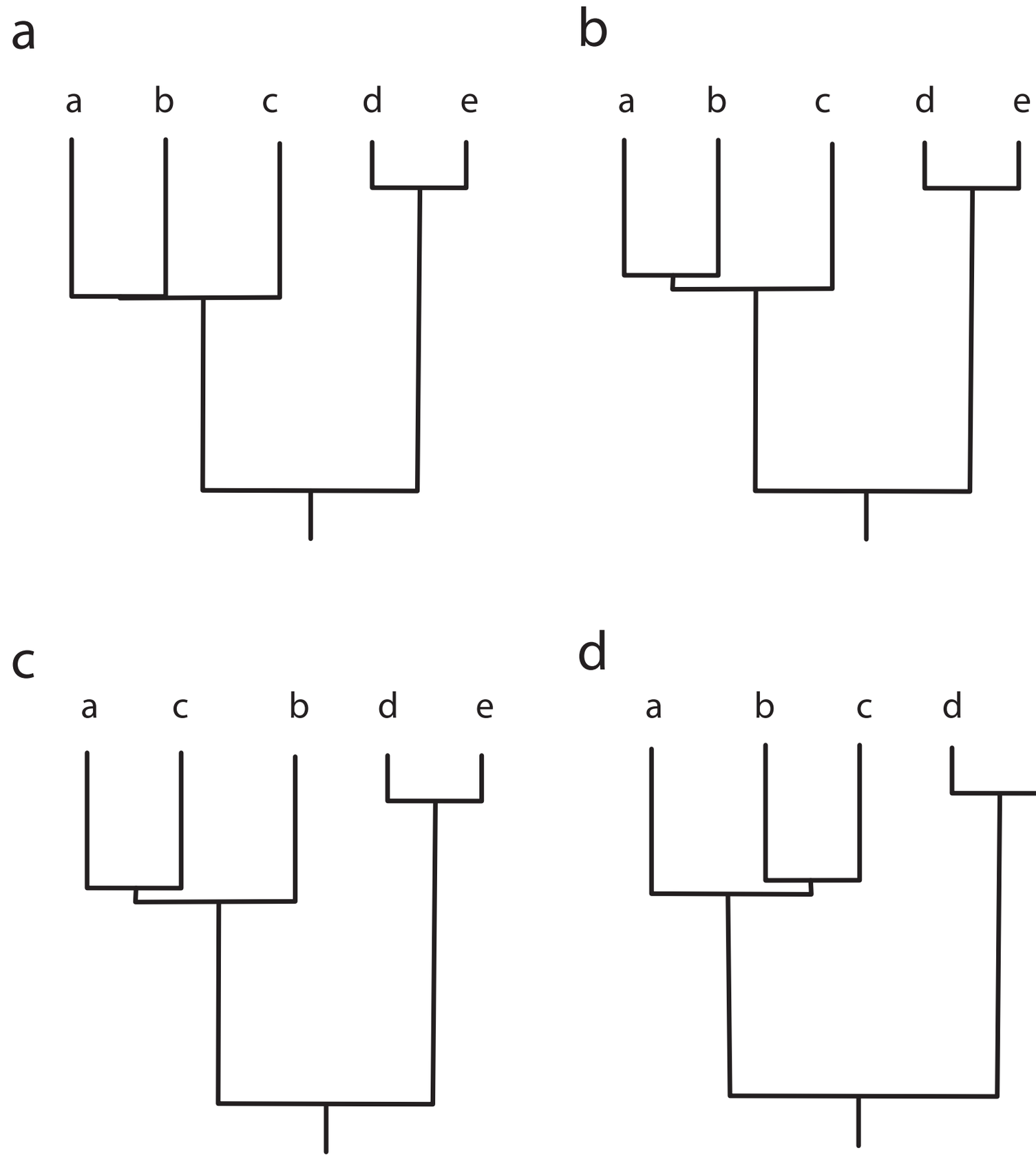


c



PHYLOGENETIC TREES




a
a b c d e
b
a b c d e
c
a c b d e
d
a b c d e

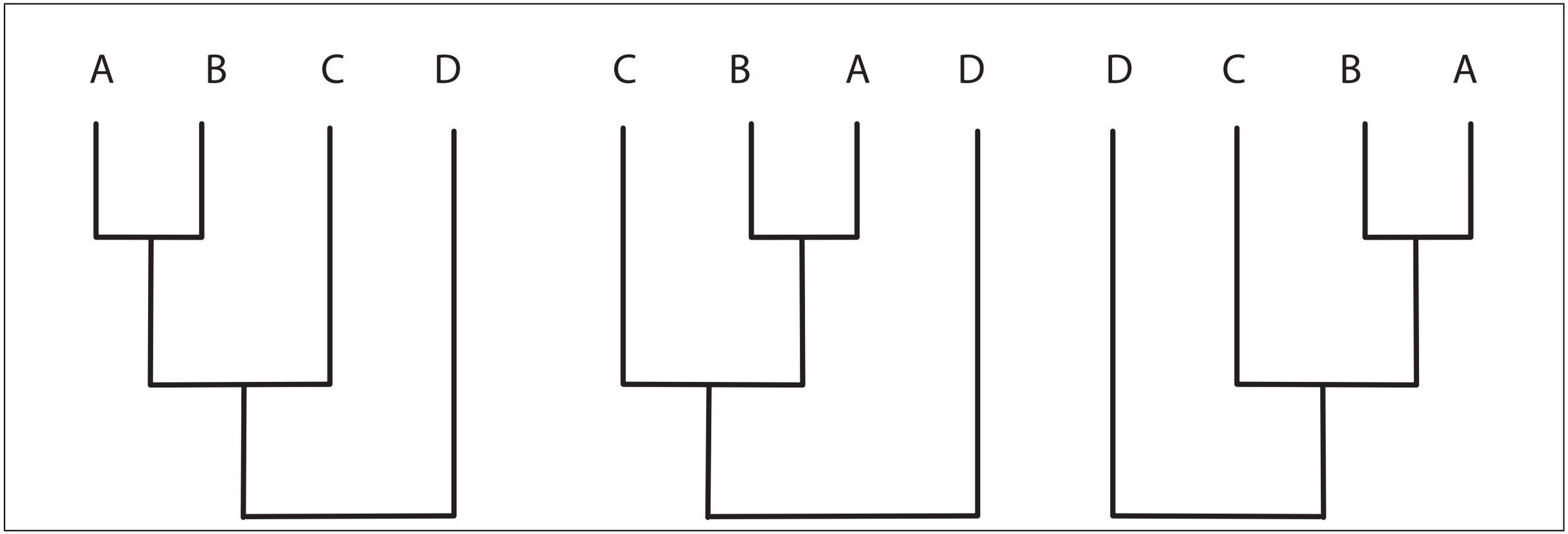


Figure 7: The same tree can be drawn in many different ways because of the rotational freedom around each node.

TABLEAU

Servant à montrer l'origine des différens animaux.

Vers.

Infusoires.
Polypes.
Radiaires.

Insectes.
Arachnides.
Crustacés.

Annelides.
Cirrhipèdes.
Mollusques.

Poissons.
Reptiles.

Oiseaux.

Monotrèmes.

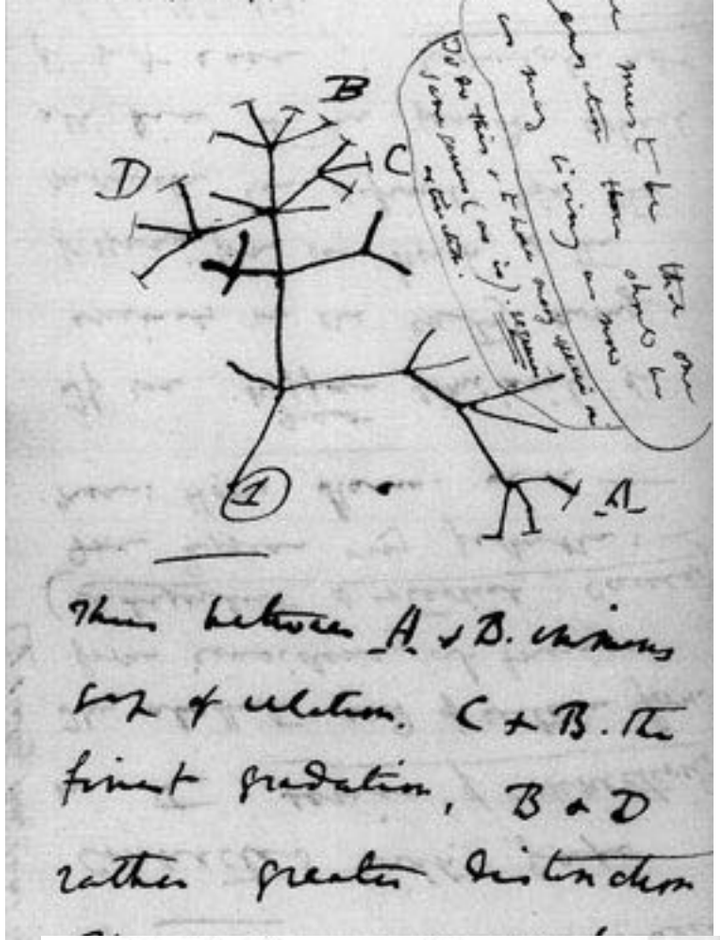
M. Amphibies.

M. Cétacés.

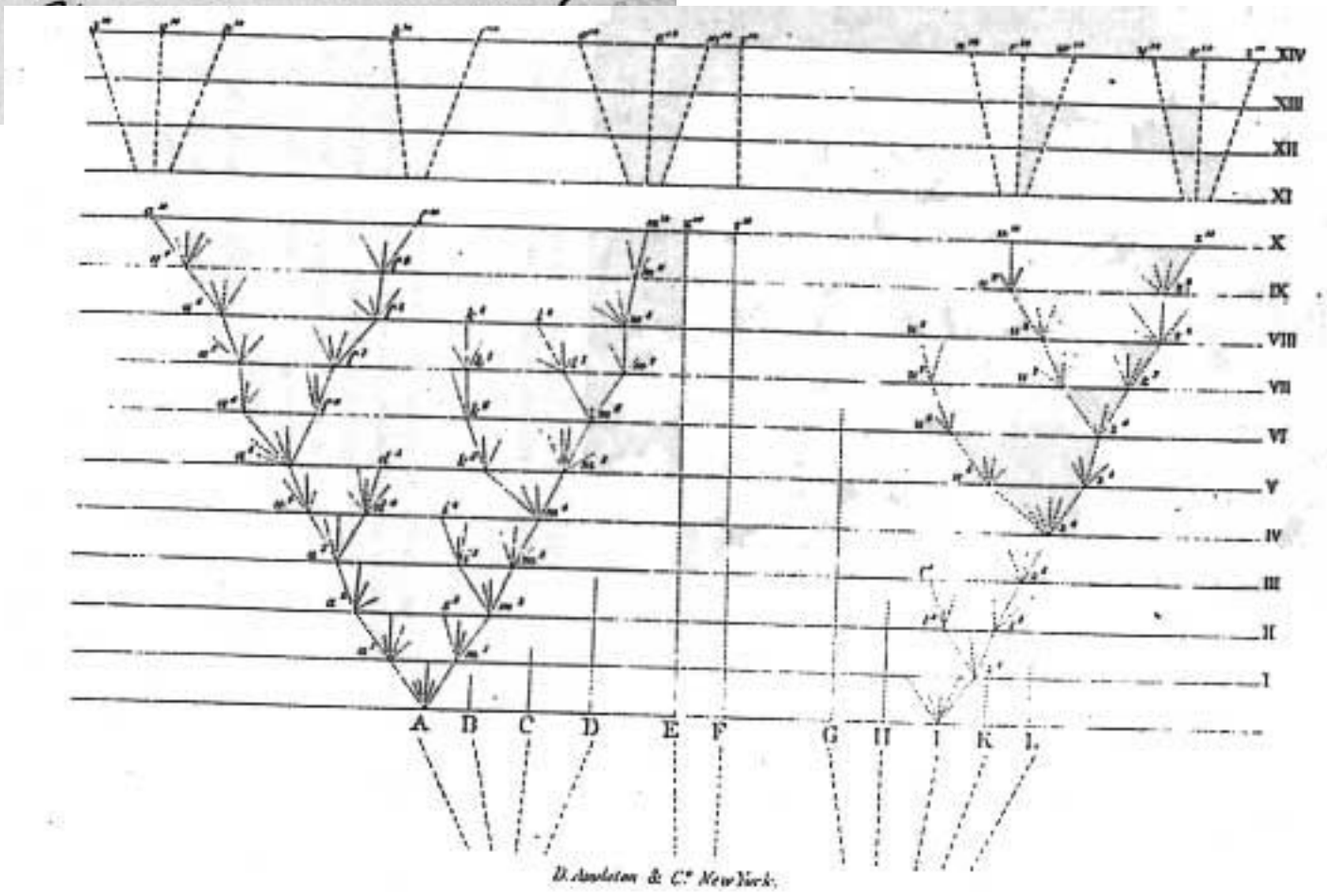
M. Ongulés.

M. Onguiculés.

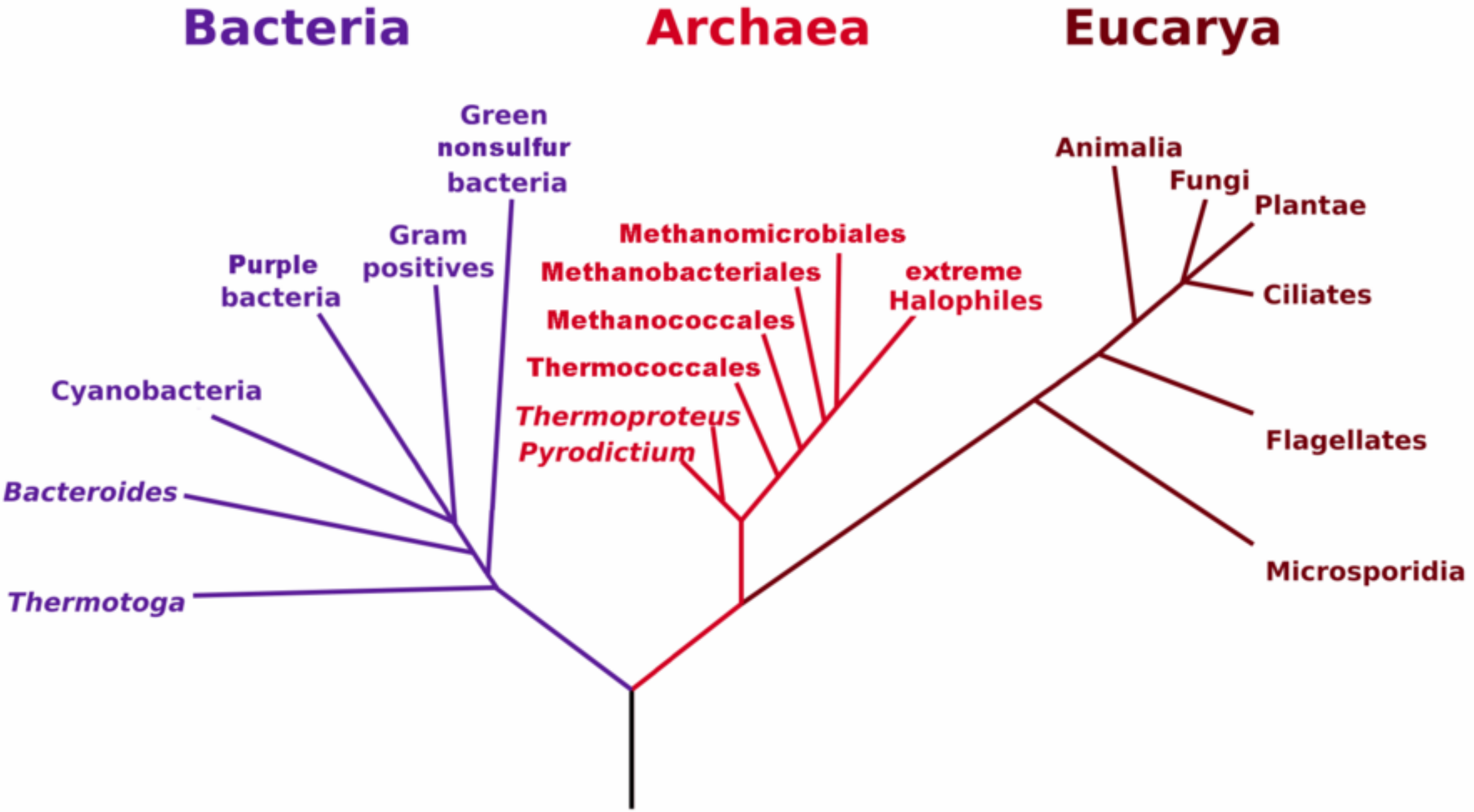
The first evolutionary tree, upside down from the modern point of view, published in Lamarck's *Philosophie zoologique* in 1809. Note the difference from the old notion of the continuous scale of nature, or chain of being. Lamarck's is a truly branching evolution. "I do not wish to say . . . that existing animals form a very simple and evenly nuanced series," he wrote, "but I say that they form a branching series irregularly graduated which has no discontinuity in its parts, or which, at least, if it is true that there are some [discontinuities] because of some lost species, has not always had such. It follows that the species which terminate each branch of the general series are related, at least on one side, to the other neighboring species which shade into them."

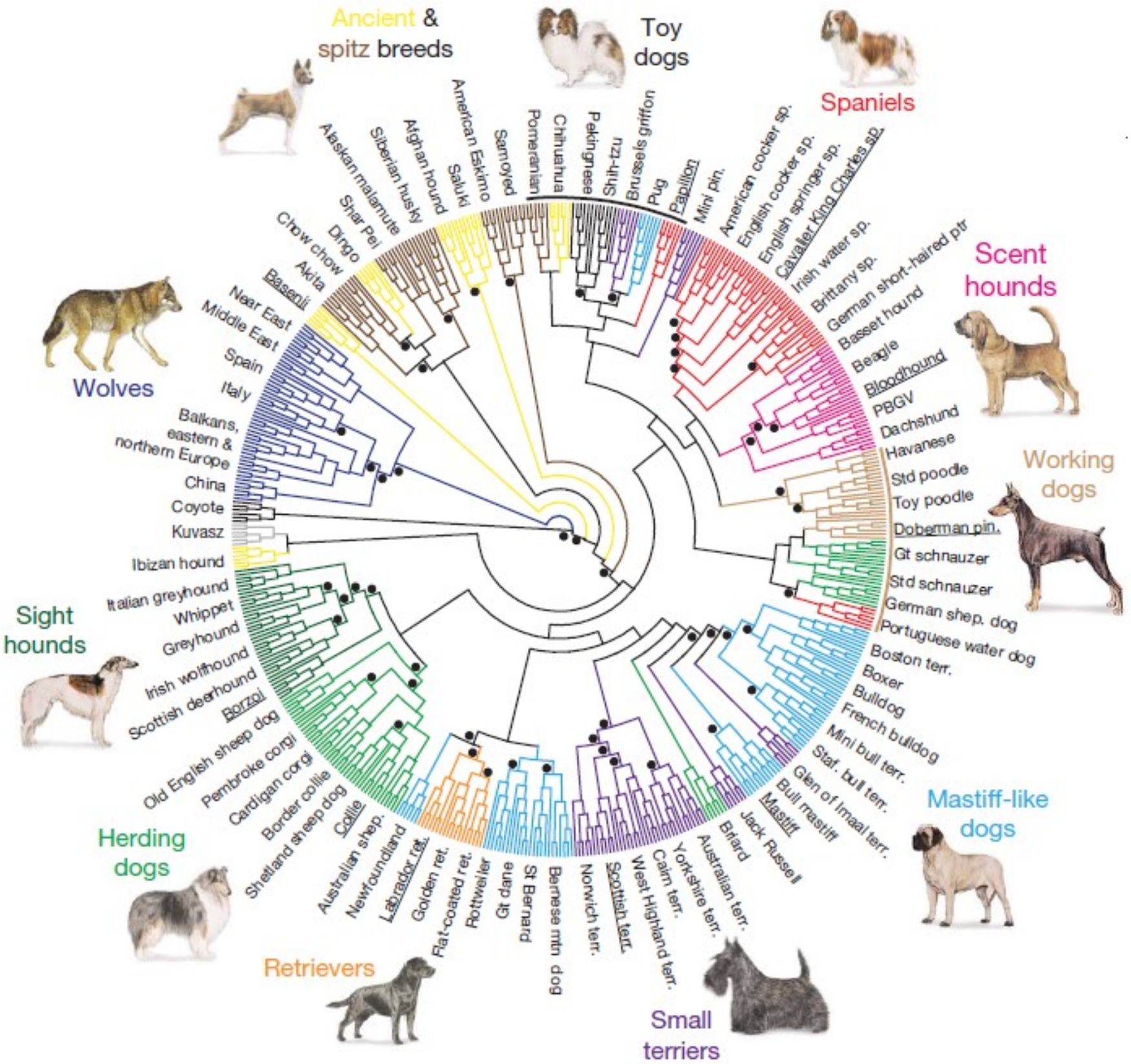


Then between A & B. causes of selection. C & B. the first gradation, B & D rather greater distinction



Phylogenetic Tree of Life




Ancient & spitz breeds
Toy dogs
Spaniels
Scent hounds
Working dogs
Mastiff-like dogs
Small terriers
Retrievers
Herding dogs
Sight hounds
Wolves
Alaskan malamute
Shar Pei
Siberian husky
Afghan hound
Saluki
American Eskimo
Samoyed
Pomeranian
Chihuahua
Pekingnese
Shih-tzu
Brussels griffon
Pug
Papillon
Mini pin.
American cocker sp.
English cocker sp.
English springer sp.
Cavalier King Charles sp.
Irish water sp.
Brittany sp.
German short-haired ptr
Basset hound
Beagle
Bloodhound
PBGV
Dachshund
Havanese
Std poodle
Toy poodle
Doberman pin.
Gt schnauzer
Std schnauzer
German shep. dog
Portuguese water dog
Boston terr.
Boxer
Bulldog
French bulldog
Mini bull terr.
Staf. bull terr.
Glen of Imaal terr.
Bull mastiff
Mastiff
Jack Russell
Briard
Australian terr.
Yorkshire terr.
Cairn terr.
West Highland terr.
Norwich terr.
Scottish terr.
Bernese mtn dog
St Bernard
Gt dane
Rottweiler
Flat-coated ret.
Golden ret.
Labrador ret.
Newfoundland
Australian shep.
Collie
Shetland sheep dog
Border collie
Cardigan corgi
Pembroke corgi
Old English sheep dog
Borzoi
Scottish deerhound
Irish wolfhound
Greyhound
Whippet
Italian greyhound
Ibizan hound
Kuvasz
Coyote
China
Balkans, eastern & northern Europe
Italy
Spain
Middle East
Near East
Basenji
Akita
Chow chow
Dingo

A

200 150 100 50 0 (Mya)

Triassic | Jurassic | Cretaceous | Paleogene | Neogene

Anura

Leiopelmatoidea

Discoglossoidea

Pipoidea

Pelobatoidea

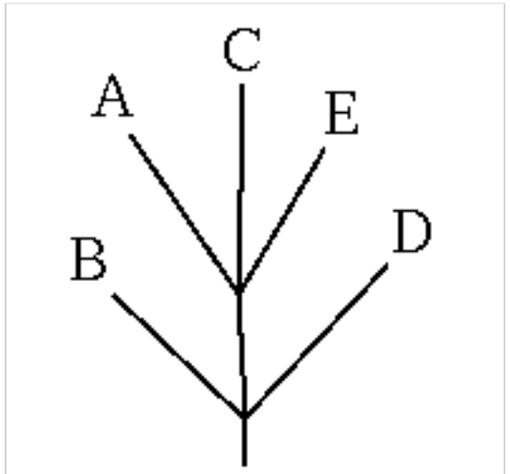
Hyloidea

Neobatrachia

Leiopelma hochstetteri — Leiopelmatidae
Ascaphus truei — Ascaphidae
Discoglossus pictus — Alytidae
Alytes obstetricans — Alytidae
Barbourula busuangensis — Bombinatoridae
Bombina orientalis — Bombinatoridae
Bombina fortinuptialis — Bombinatoridae
Rhinophrynus dorsalis — Rhinophrynidae
Pipa pipa — Pipidae
Pipa parva — Pipidae
Pseudhymenochirus merlini — Pipidae
Hymenochirus boettgeri — Pipidae
Xenopus laevis — Pipidae
Xenopus epitropicalis — Pipidae
Scaphiopus couchii — Scaphiopodidae
Spea multiplicata — Scaphiopodidae
Spea intermontana — Scaphiopodidae
Pelodytes ibericus — Pelodytidae
Pelobates syriacus — Pelobatidae
Brachytarsophrys feae — Megophryidae
Xenophrys wuliangshanensis — Megophryidae
Ophryophryne microstoma — Megophryidae
Leptolalax alpinus — Megophryidae
Scutiger gongshanensis — Megophryidae
Leptobrachium chapaense — Megophryidae
Oreolalax jingdongensis — Megophryidae
Heleophryne purcelli — Heleophrynidae
Calyptocephalella gayi — Calyptocephalellidae
Myobatrachidae — Myobatrachidae
Crinia signifera — Myobatrachidae
Limnodynastes salmini — Myobatrachidae
Insuetophrynus acarpicus — Rhinodermatidae
Rhinoderma darwinii — Rhinodermatidae
Alsodes gargola — Alsodidae
Eupsophus calcaratus — Alsodidae
Ateleognathus reverberi — Batrachylidae
Batrachyla taeniata — Batrachylidae
Batrachyla leptopus — Batrachylidae
Batrachophrynus macrostomus — Telmatobiidae
Telmatobius vellardi — Telmatobiidae
Lepidobatrachus sp. — Ceratophryidae
Ceratophrys cornuta — Ceratophryidae
Cryptobatrachus boulengeri — Hemiphractidae
Gastrotheca pseustes — Hemiphractidae
Gastrotheca weinlandii — Hemiphractidae
Nyctimystes kubori — Hylidae
Phyllomedusa tomopterna — Hylidae
Agalychnis callidryas — Hylidae
Agalychnis lemur — Hylidae
Hylomantis lindae — Hylidae
Aplastodiscus perviridis — Hylidae
Hypsiboas fasciatus — Hylidae
Scinax ruber — Hylidae
Osteocephalus taurinus — Hylidae
Hyla chinensis — Hylidae
Acris crepitans — Hylidae
Dendropsophus parviceps — Hylidae
Pseudis paradoxa — Hylidae
Allobates femoralis — Dendrobatidae
Hylaxelus jacobuspeteri — Dendrobatidae
Ranitomeya imitator — Dendrobatidae
Fleutherodactylus planirostris — Eleutherodactylidae
Craugastor augusti — Craugastoridae
Craugastor fitzingeri — Craugastoridae
Pristimantis thymelensis — Strabomantidae
Barycholos ulcerus — Strabomantidae
Strabomantis sulcatus — Strabomantidae
Hypodactylus brunneus — Strabomantidae
Proceratophrys boiei — Odontophrynidae
Odontophrynus occidentalis — Odontophrynidae
Leptodactylus albilabris — Leptodactylidae
Lithodytes lineatus — Leptodactylidae
Physalaemus pustulosus — Leptodactylidae
Physalaemus cuvieri — Leptodactylidae
Pleurodema somuncurensis — Leptodactylidae
Pleurodema thaul — Leptodactylidae
Melanophryniscus stelzneri — Bufonidae
Amazophrynella minuta — Bufonidae
Peltophryne peltocephala — Bufonidae
Schismaderma carens — Bufonidae

Archaeobatrachia

NEWICK TREE FORMAT



The Newick Standard for representing trees in computer-readable form makes use of the correspondence between trees and nested parentheses, noticed in 1857 by the famous English mathematician Arthur Cayley. If we have this rooted tree on the left then in the tree file it is represented by the following sequence of printable characters:

(B,(A,C,E),D);

The tree ends with a semicolon. The bottommost node in this tree is an interior node, not a tip. Interior nodes are represented by a pair of matched parentheses. Between them are representations of the nodes that are immediately descended from that node, separated by commas. In the above tree, the immediate descendants are B, another interior node, and D. The other interior node is represented by a pair of parentheses, enclosing representations of its immediate descendants, A, C, and E. In our example these happen to be tips, but in general they could also be interior nodes and the result would be further nestings of parentheses, to any level.
Tips are represented by their names. A name can be any string of printable characters except blanks, colons, semicolons, parentheses, and square brackets.

NEWICK TREE FORMAT

Because you may want to include a blank in a name, it is assumed that an underscore character ("_") stands for a blank; any of these in a name will be converted to a blank when it is read in. Any name may also be empty: a tree like
(,(,,),);
is allowed. Trees can be multifurcating at any level.

Branch lengths can be incorporated into a tree by putting a real number, with or without decimal point, after a node and preceded by a colon. This represents the length of the branch immediately below that node. Thus the above tree might have lengths represented as:

(B:6.0,(A:5.0,C:3.0,E:4.0):5.0,D:11.0);

NEWICK TREE FORMAT

The tree starts on the first line of the file, and can continue to subsequent lines. It is best to proceed to a new line, if at all, immediately after a comma. Blanks can be inserted at any point except in the middle of a species name or a branch length.
The above description is actually of a subset of the Newick Standard. For example, interior nodes can have names in that standard. These names follow the right parenthesis for that interior node, as in this example:

(B:6.0,(A:5.0,C:3.0,E:4.0)Ancestor1:5.0,D:11.0);

Examples
To help you understand this tree representation, here are some trees in the above form:

((raccoon:19.19959,bear:6.80041):0.84600,((sea_lion:11.99700, seal:12.00300):7.52973,((monkey:100.85930,cat:47.14069):20.59201, weasel:18.87953):2.09460):3.87382,dog:25.46154);

(Bovine:0.69395,(Gibbon:0.36079,(Orang:0.33636,(Gorilla:0.17147,(Chimp:0.19268, Human:0.11927):0.08386):0.06124):0.15057):0.54939,Mouse:1.21460):0.10;

(Bovine:0.69395,(Hylobates:0.36079,(Pongo:0.33636,(G._Gorilla:0.17147, (P._paniscus:0.19268,H._sapiens:0.11927):0.08386):0.06124):0.15057):0.54939, Rodent:1.21460);
A;

((A,B),(C,D));

(Alpha,Beta,Gamma,Delta,,Epsilon,,,);

(Non-)Uniqueness

The Newick Standard does not make a unique representation of a tree, for two reasons. First, the left-right order of descendants of a node affects the representation, even though it is biologically uninteresting. Thus, to a biologist

```
(A,(B,C),D);
```

is the same tree as

```
(A,(C,B),D);
```

which is in turn the same tree as

```
(D,(C,B),A);
```

and that is the same tree as

```
(D,A,(C,B));
```

and

```
((C,B),A,D);
```

Rooted and unrooted trees

In addition, the standard is representing a rooted tree. For many biological purposes we may not be able to infer the position of the root. We would like to have a representation of an unrooted tree when describing inferences in such cases. Here the convention is simply to arbitrarily root the tree and report the resulting rooted tree. Thus

```
(B,(A,D),C);
```

would be the same unrooted tree as

```
(A,(B,C),D);
```

and as

```
((A,D),(C,B));
```

NEWICK TREE FORMAT

The Newick Standard was adopted 26 June 1986 by an informal committee meeting convened by Joe Felsenstein during the Society for the Study of Evolution meetings in Durham, New Hampshire and consisting of James Archie, William H.E. Day, Wayne Maddison, Christopher Meacham, F. James Rohlf, David Swofford, and myself. (The committee was not an activity of the SSE nor endorsed by it). The reason for the name is that the second and final session of the committee met at Newick's restaurant in Dover, New Hampshire, and we enjoyed the meal of lobsters. The tree representation was a generalization of one developed by Christopher Meacham in 1984 for the tree plotting programs that he wrote for the PHYLIP package while visiting Seattle. His visit was a sabbatical leave from the University of Georgia, which thus indirectly partly funded that work.

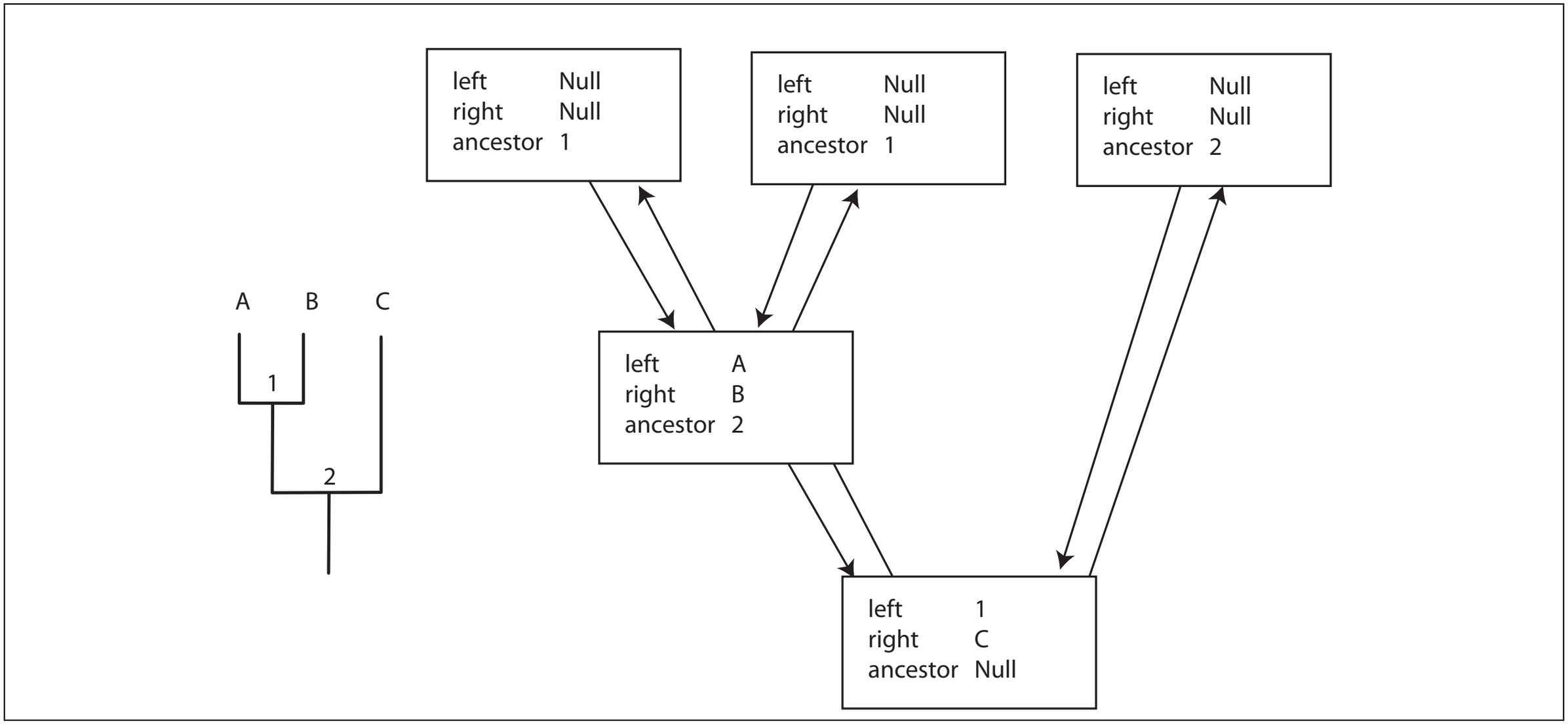


Figure 8: The Binary Tree Data Model (BTM).

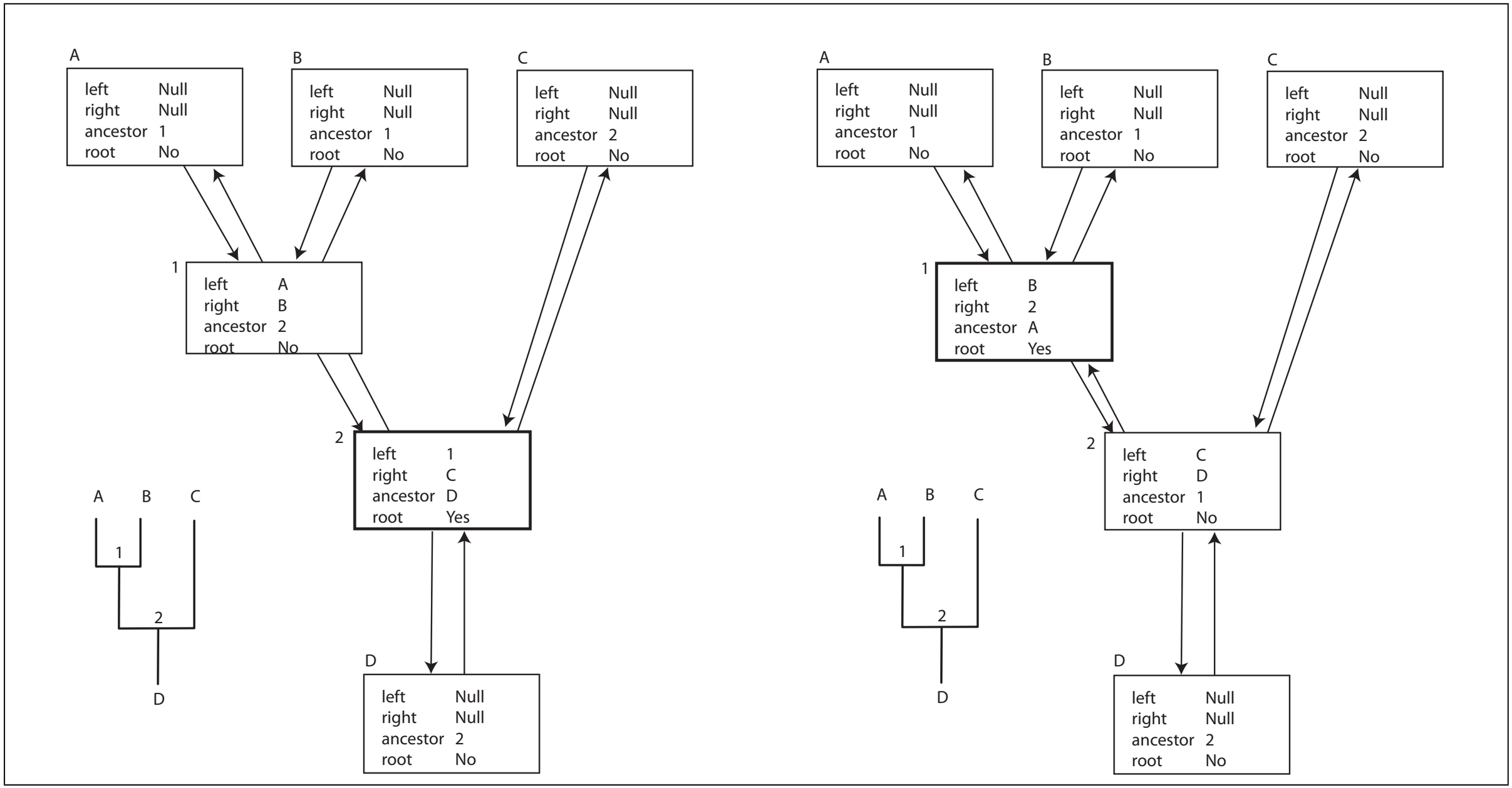


Figure 9: A modification of BTM to accommodate unrooted trees. The two trees have different roots, the pointers change in all interior nodes.

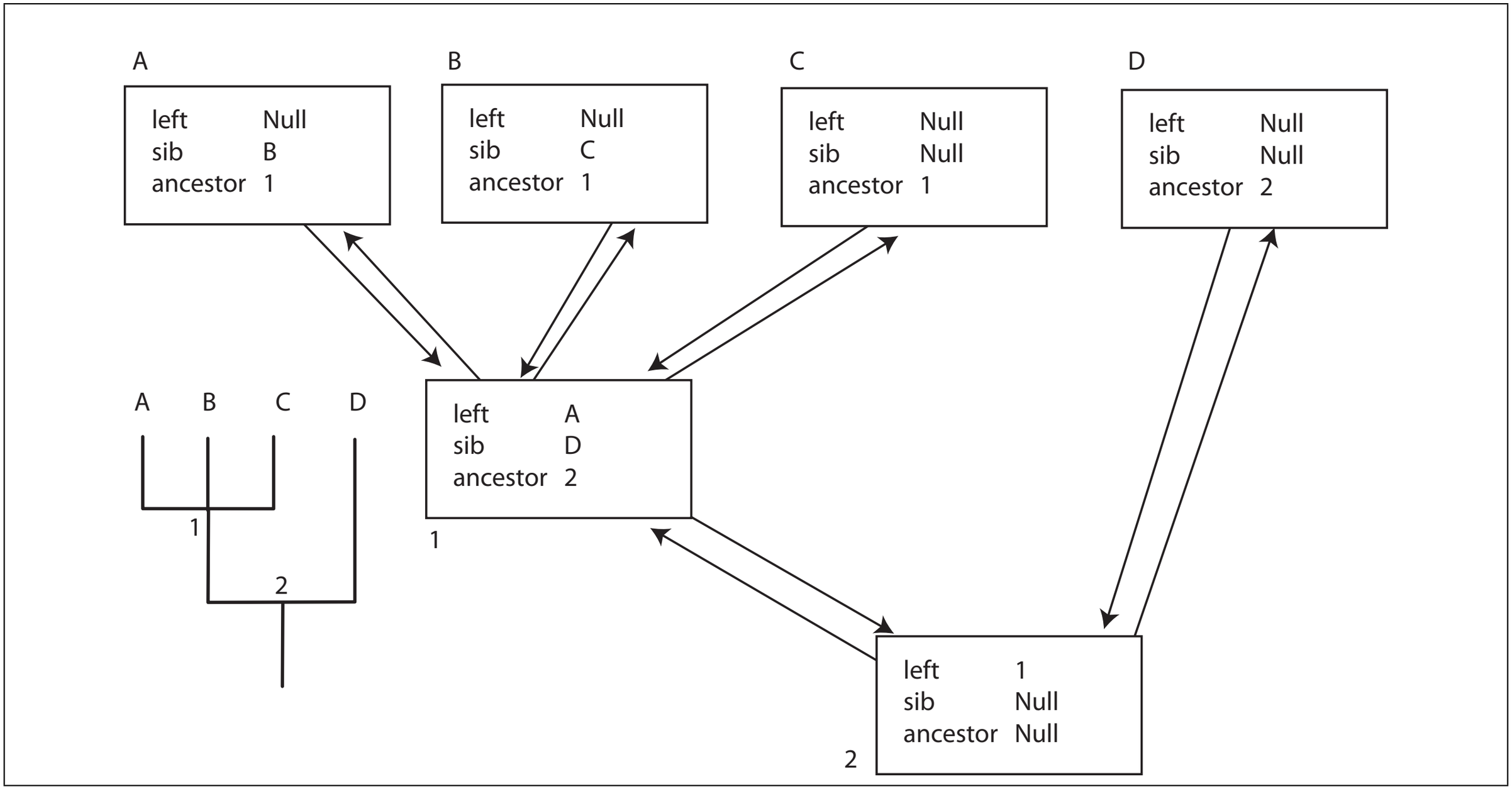


Figure 10: The polytomous tree data model (PTM).

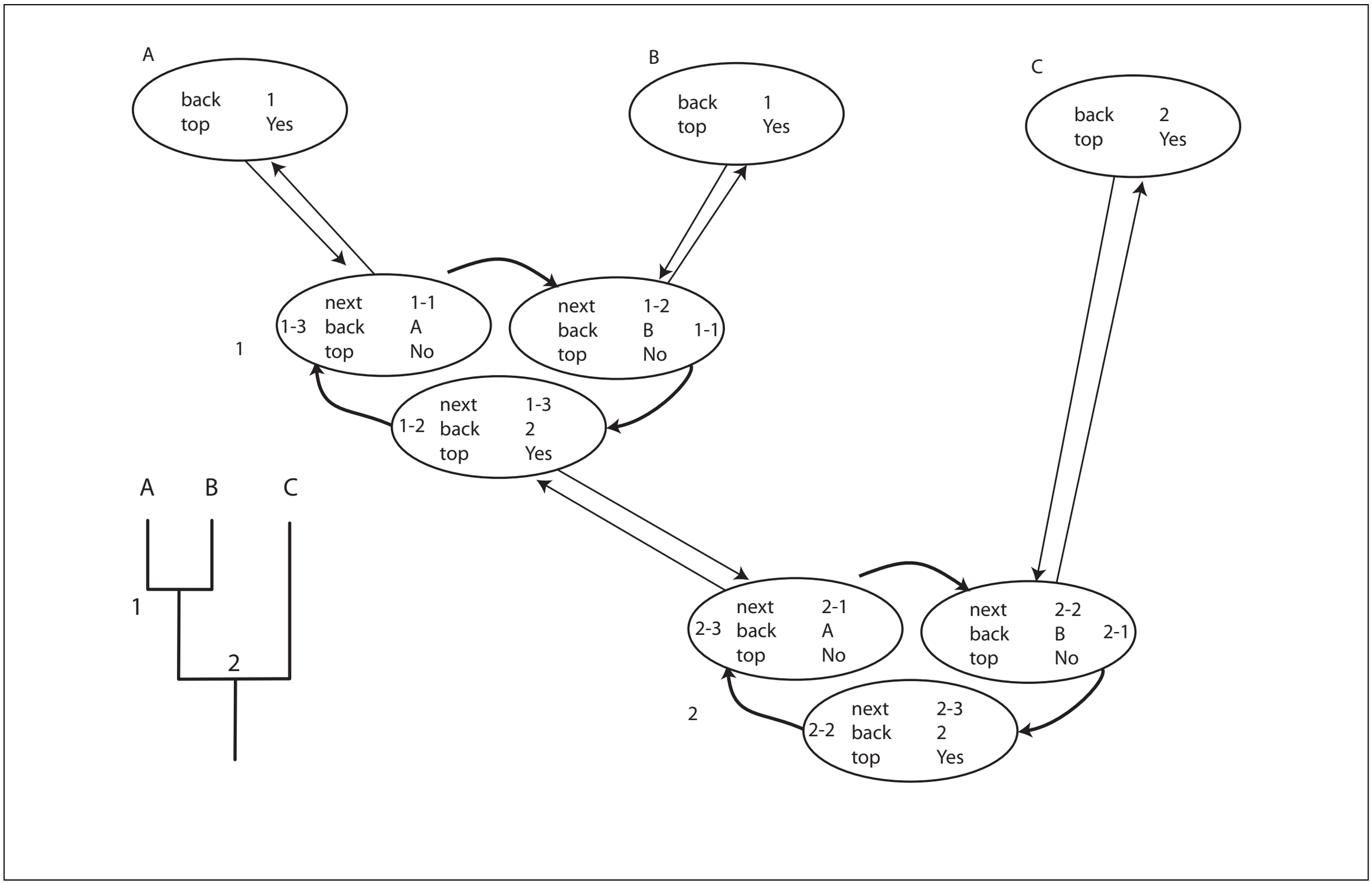


Figure 11: The Felsenstein tree data model (FTM). Each node is broken up into nodelets, this facilitates greatly rerooting and polytomous trees.

Algorithm 1 Postorder traversal algorithm

Traverse left descendant of p

Traverse right descendant of p

Carry out $f(p)$ on p

Algorithm 2 Preorder traversal algorithm

Carry out $f(p)$ on p

Traverse left descendant of p

Traverse right descendant of p

Algorithm 3 Recursive algorithm for printing a tree (BTM unordered or ordered)

```
if p is not tip then
    Print '(' and left subtree of p
    Print ',' and right subtree of p
    if p is 'root' of unrooted tree then
        Print ',' and ancestor subtree of p
    end if
    Print ')'
end if
Print name of p (if any)
Print ':' and branch length of p (if any)
```