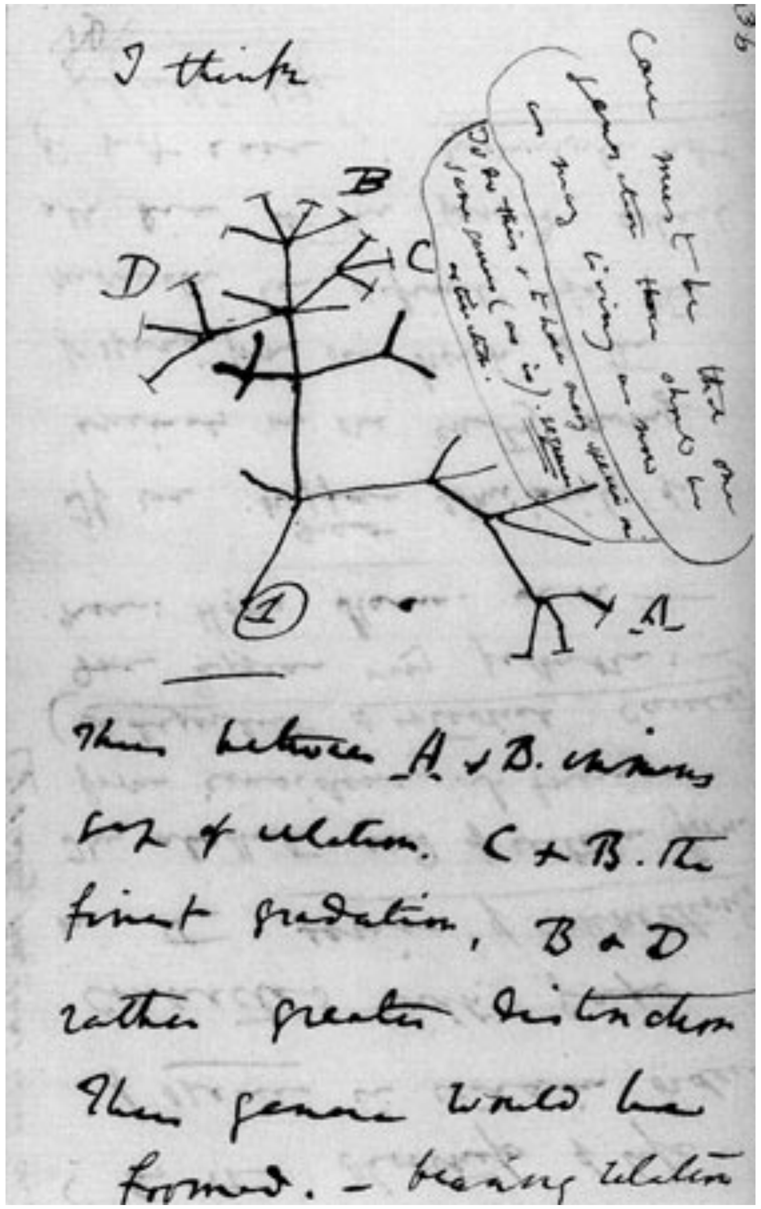


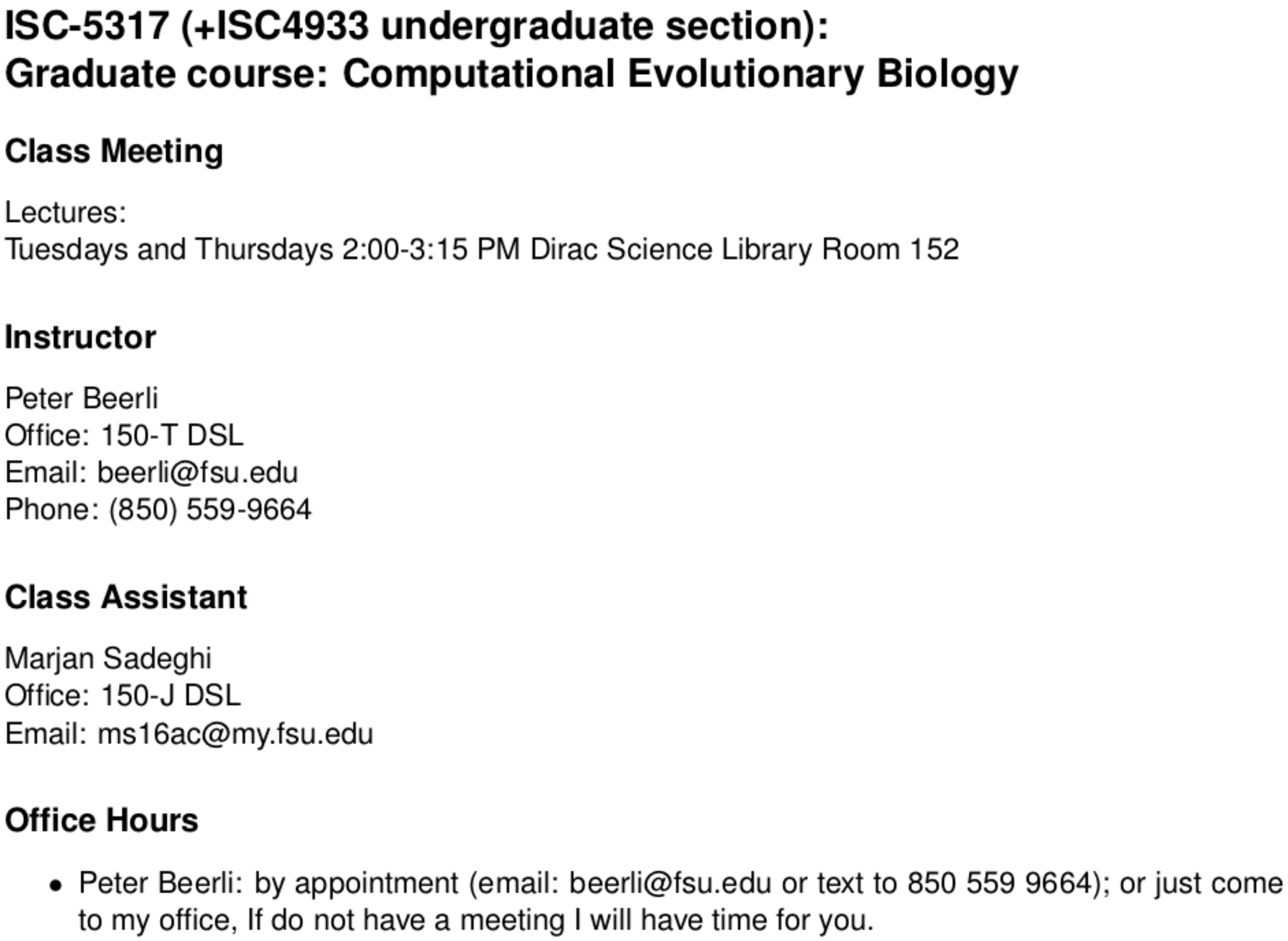
I think

Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is) requires extinction.

Thus between A & B. immense gap of relation. C & B. the finest gradation, B & D rather greater distinction Thus genera would be formed. — bearing relation

ISC 5317 / ISC 4933

CEB COMPUTATIONAL EVOLUTIONARY BIOLOGY



ISC-5317 (+ISC4933 undergraduate section): Graduate course: Computational Evolutionary Biology

Class Meeting

Lectures:
Tuesdays and Thursdays 2:00-3:15 PM Dirac Science Library Room 152

Instructor

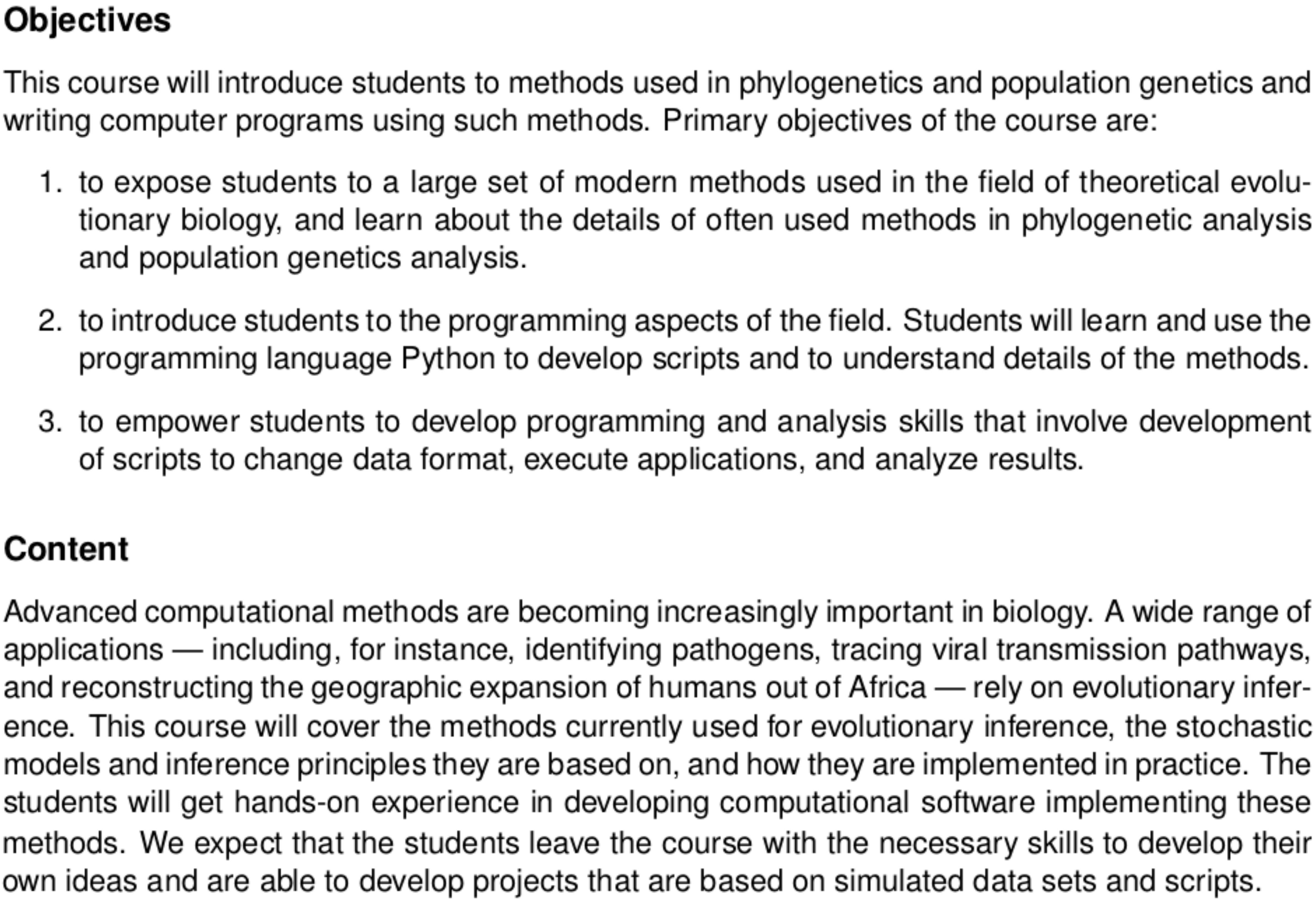
Peter Beerli
Office: 150-T DSL
Email: beerli@fsu.edu
Phone: (850) 559-9664

Class Assistant

Marjan Sadeghi
Office: 150-J DSL
Email: ms16ac@my.fsu.edu

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.



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:

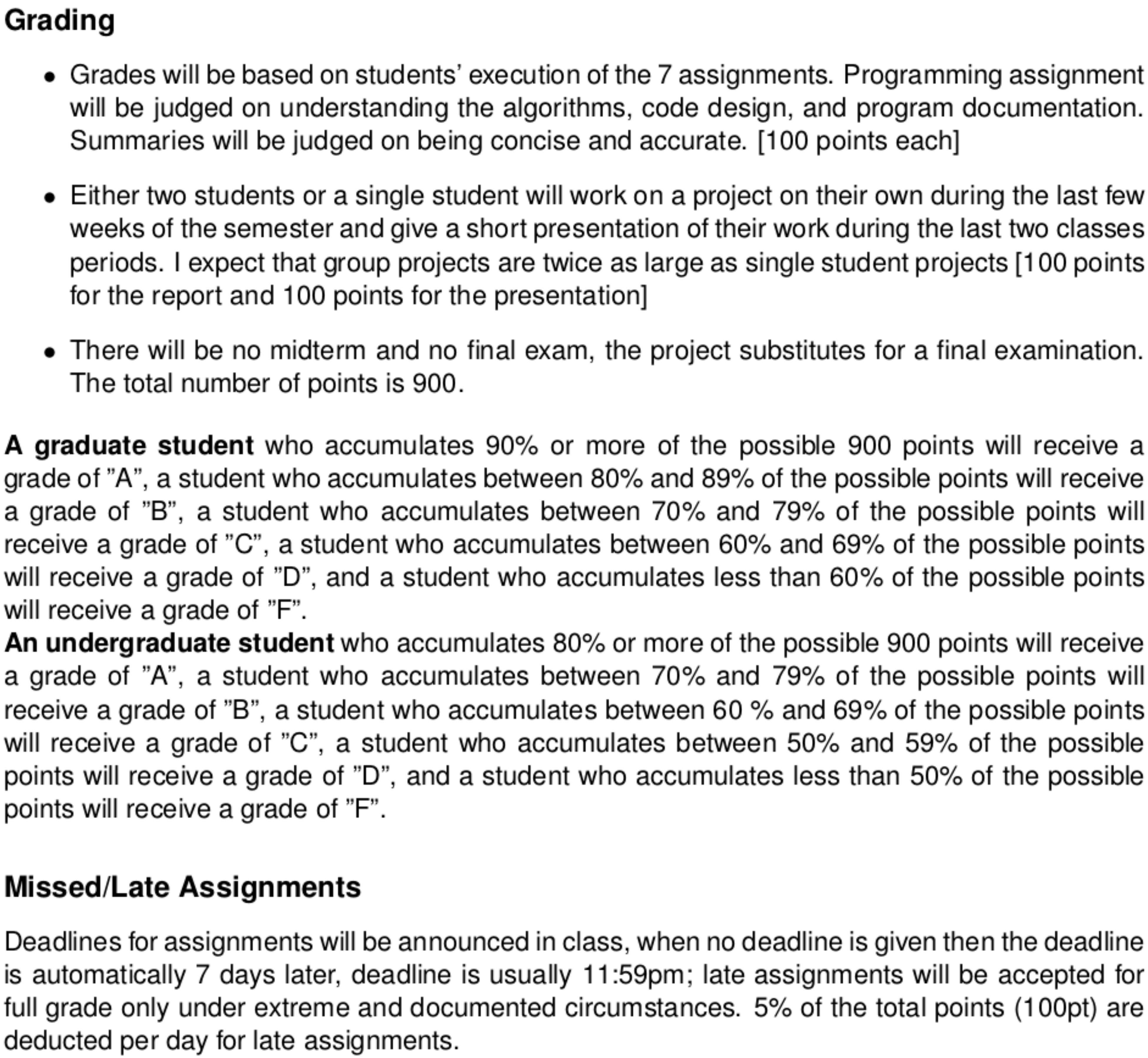
1. to expose students to a large set of modern methods used in the field of theoretical evolutionary biology, and learn about the details of often used methods in phylogenetic analysis and population genetics analysis.

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.



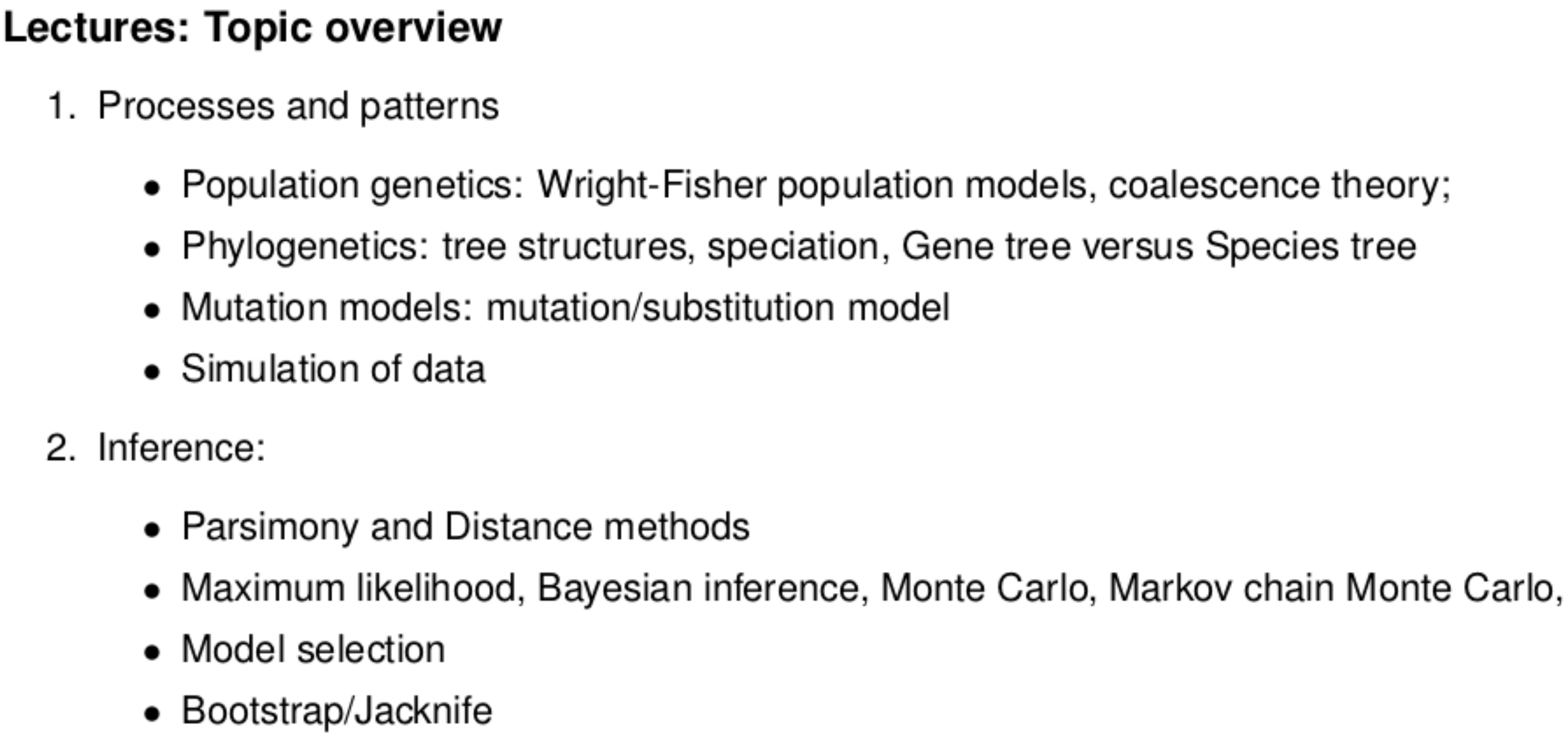
Grading

- Grades will be based on students’ execution of the 7 assignments. Programming assignment will be judged on understanding the algorithms, code design, and program documentation. Summaries will be judged on being concise and accurate. [100 points each]
- Either two students or a single student will work on a project on their own during the last few weeks of the semester and give a short presentation of their work during the last 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]
- There will be no midterm and no final exam, the project substitutes for a final examination. The total number of points is 900.

A graduate student who accumulates 90% or more of the possible 900 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”.
An undergraduate student who accumulates 80% or more of the possible 900 points will receive a grade of ”A”, a student who accumulates between 70% and 79% of the possible points will receive a grade of ”B”, a student who accumulates between 60 % and 69% of the possible points will receive a grade of ”C”, a student who accumulates between 50% and 59% of the possible points will receive a grade of ”D”, and a student who accumulates less than 50% of the possible points will receive a grade of ”F”.

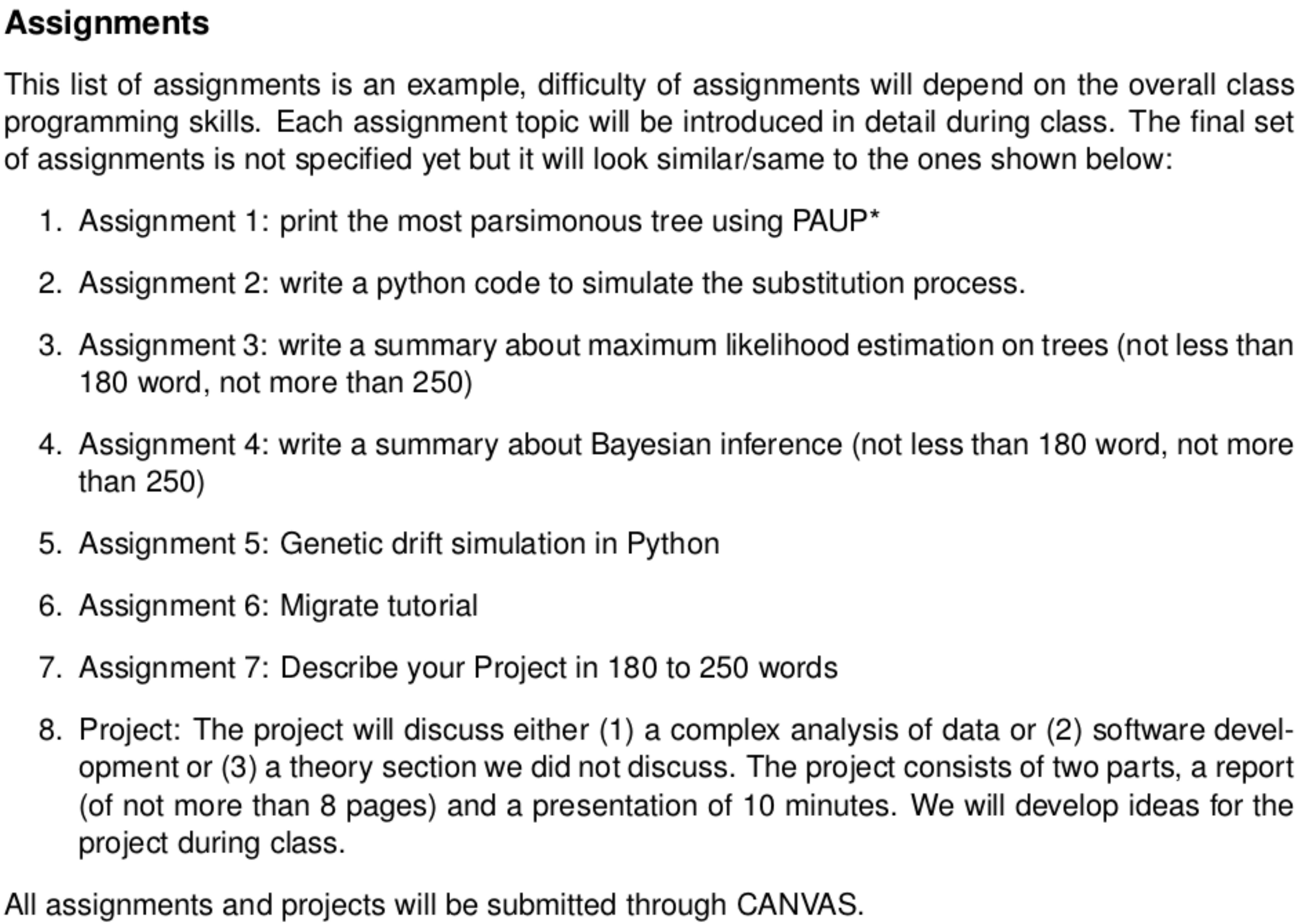
Missed/Late Assignments

Deadlines for assignments will be announced in class, when no deadline is given then the deadline is automatically 7 days later, deadline is usually 11:59pm; late assignments will be accepted for full grade only under extreme and documented circumstances. 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



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/same to the ones shown below:

1. Assignment 1: print the most parsimonous tree using PAUP*

2. Assignment 2: write a python code to simulate the substitution process.

3. Assignment 3: write a summary about maximum likelihood estimation on trees (not less than 180 word, not more than 250)

4. Assignment 4: write a summary about Bayesian inference (not less than 180 word, not more than 250)

5. Assignment 5: Genetic drift simulation in Python

6. Assignment 6: Migrate tutorial

7. Assignment 7: Describe your Project in 180 to 250 words

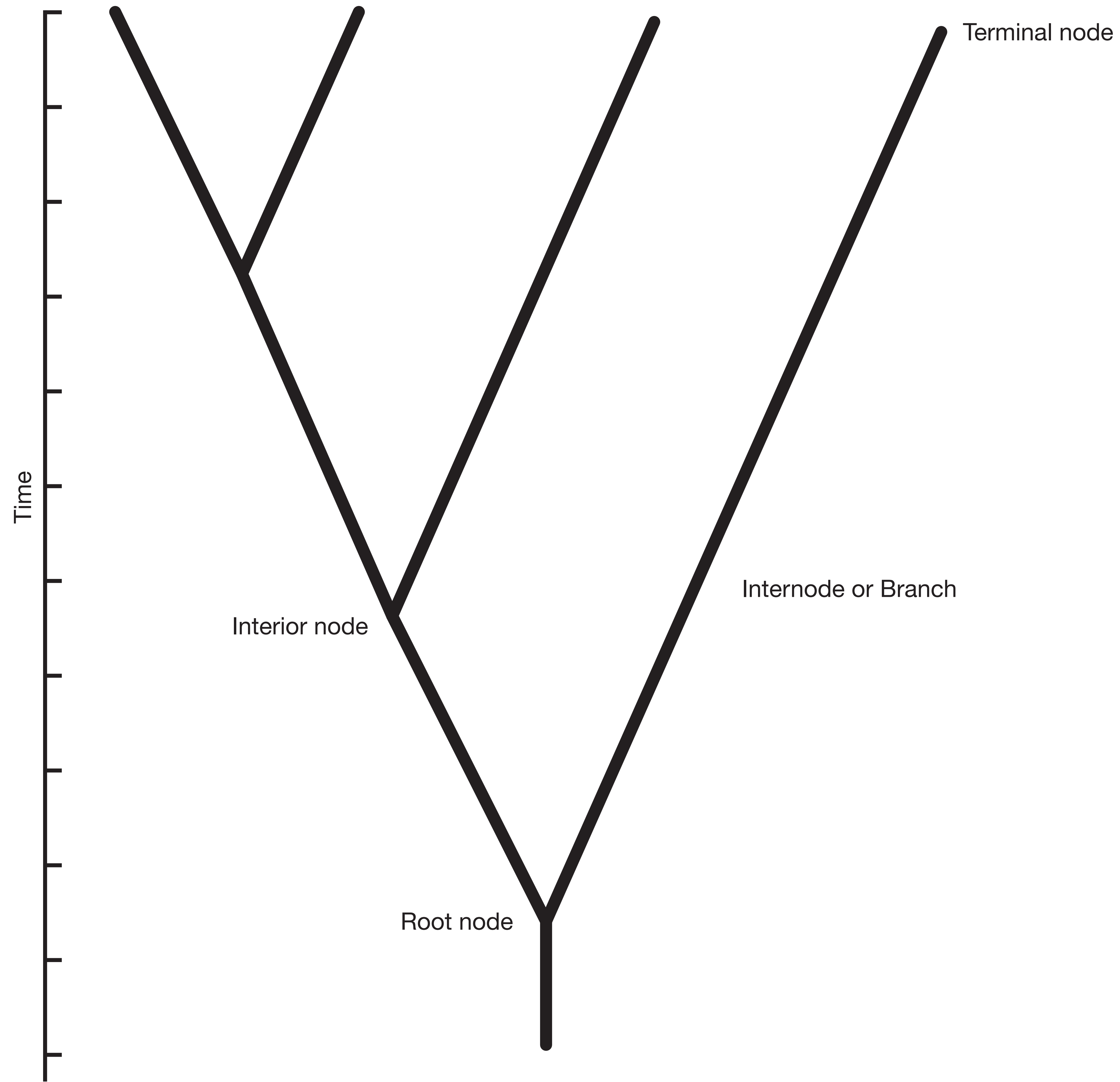
8. Project: The project will discuss either (1) a complex analysis of data or (2) software development or (3) a theory section we did not discuss. The project consists of two parts, a report (of not more than 8 pages) and a presentation of 10 minutes. We will develop ideas for the project during class.

All assignments and projects will be submitted through CANVAS.

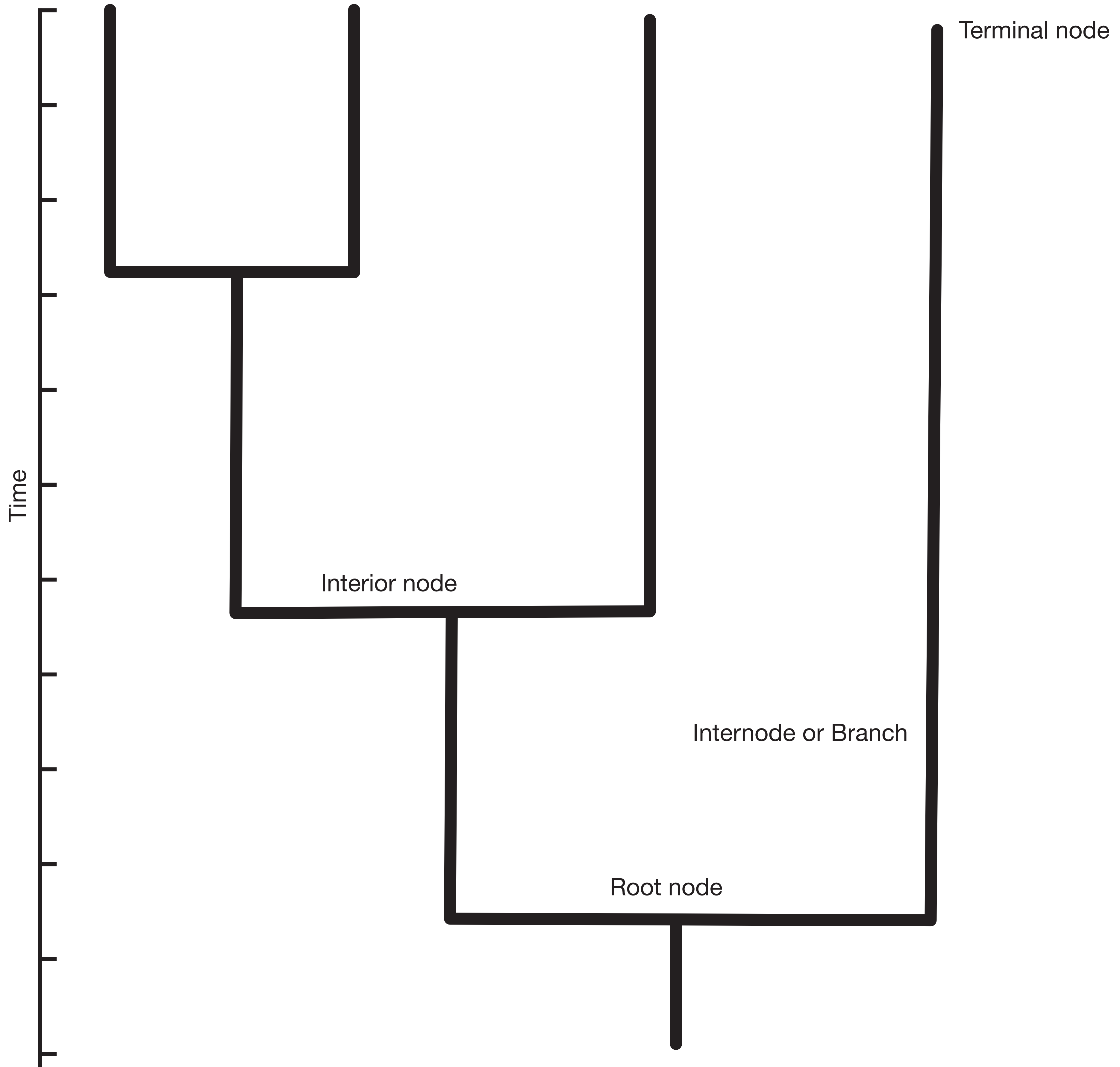
Lecture Schedule

1. Introduction. Trees and tree representation (Aug. 27)
2. Python and trees [Self study: Python tutorial] (Aug. 29)
3. Selfstudy: Parsimony (Sep 3)
4. Selfstudy: Install PAUP* [Assignment 1] (Sep 5)
5. Number of trees; Searching for the best tree(s) (Sep 10)
6. Substitution models and distance measures(Sep 12)
7. Substitution models exercise (Sep 17) [Assignment 2]
8. Paul Lewis: Maximum likelihood, substitution model and trees (Sep 19)
9. Paul Lewis: Tree likelihood and rate heterogeneity (Sep 24) [Assignment 3]
10. Paul Lewis: Bayesian inference and Markov Chain Monte Carlo (Sep 26)
11. Paul Lewis: Bayesian inference on Trees (Oct 1) [Assignment 4]
12. Question and Answer session about Paul Lewis lecture (Oct 3)
13. Population genetics introduction (Oct 8)
14. Population simulation in Python [Assignment 5] (Oct 10)
15. The coalescent (Oct 15)
16. Coalescent simulation and extensions to the coalescent (Oct 17)
17. Coalescent simulation in Python [Assignment 6] (Oct 22)
18. Gene tree vs Species tree (Oct 24)
19. SVD quartets and other PAUP* evaluations (Oct 29)
20. Analysis of admixture, networks (Nov 5)
21. Model Selection (Nov 7) [Assignment 7]
22. Model Selection exercise using Migrate (Nov 12)
23. Bootstrap/Jacknife with PAUP* (Nov 14)
24. Project (Nov 19)
25. Project (Nov 21)
26. Project (Nov 26)
27. Thanksgiving break
28. Presentation (Dec 3)

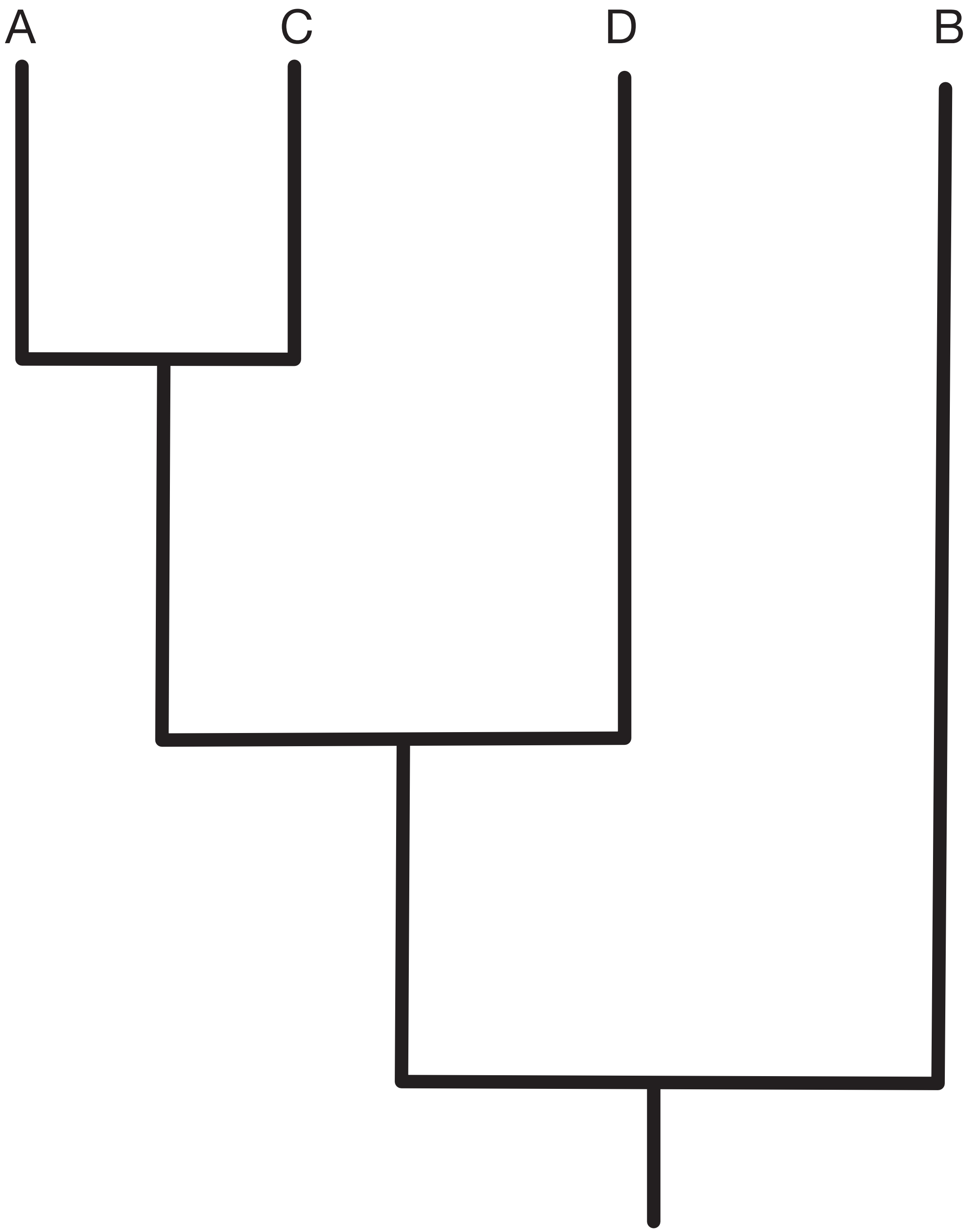
PHYLOGENETIC TREES

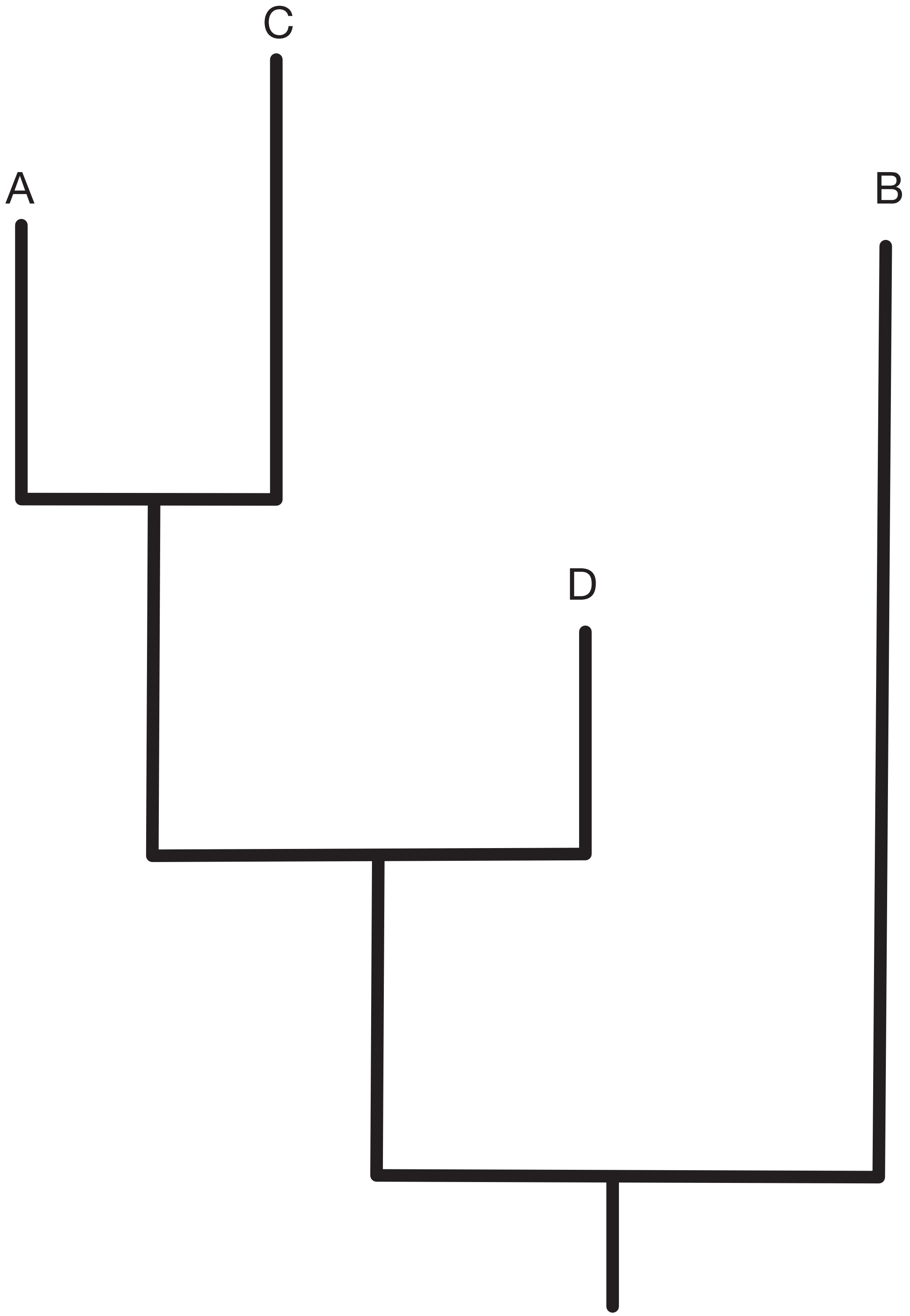


PHYLOGENETIC TREES

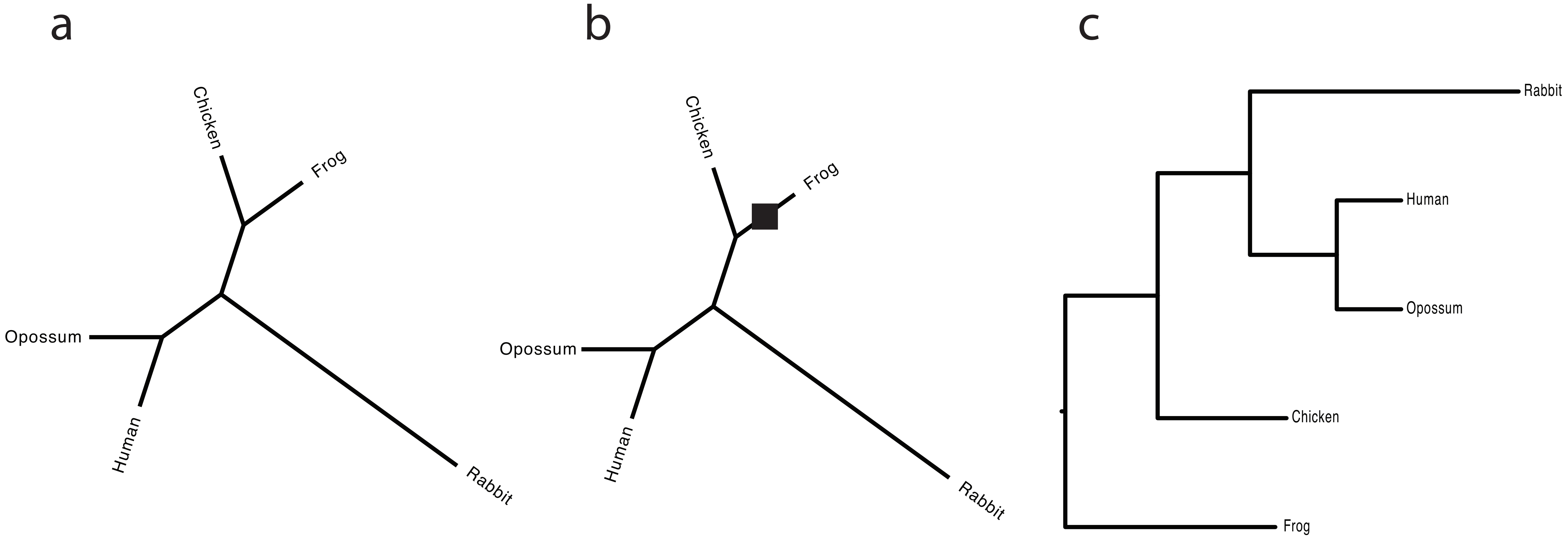


PHYLOGENETIC TREES

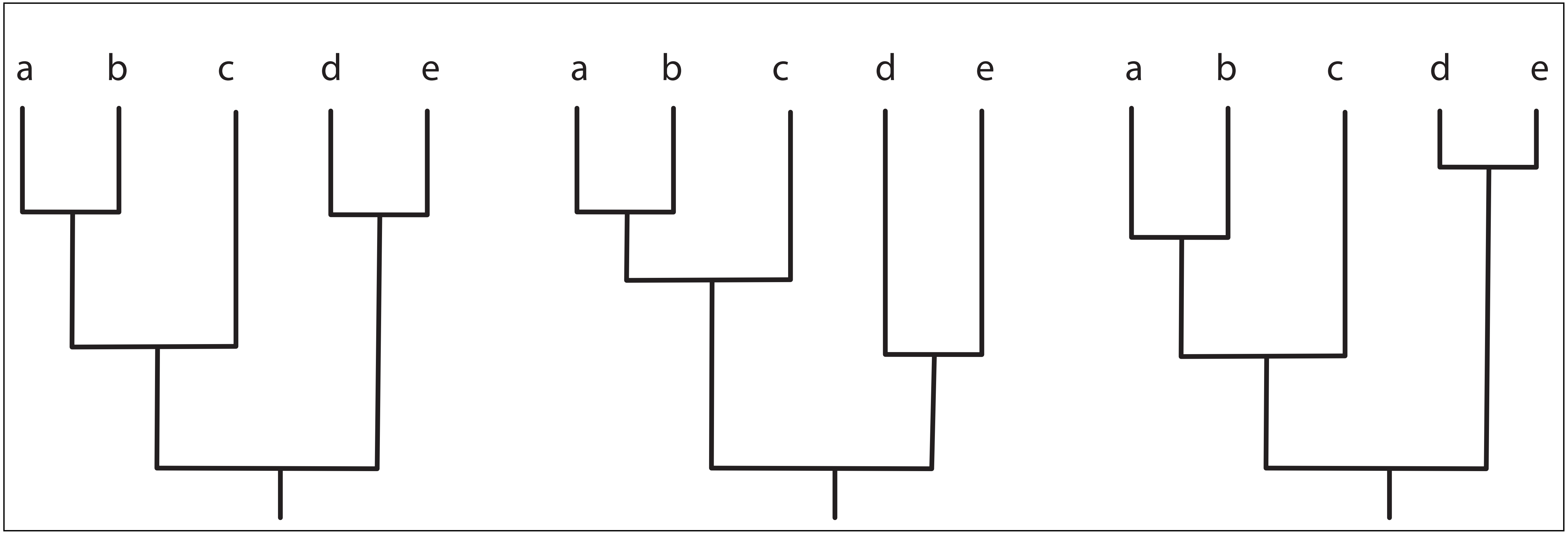




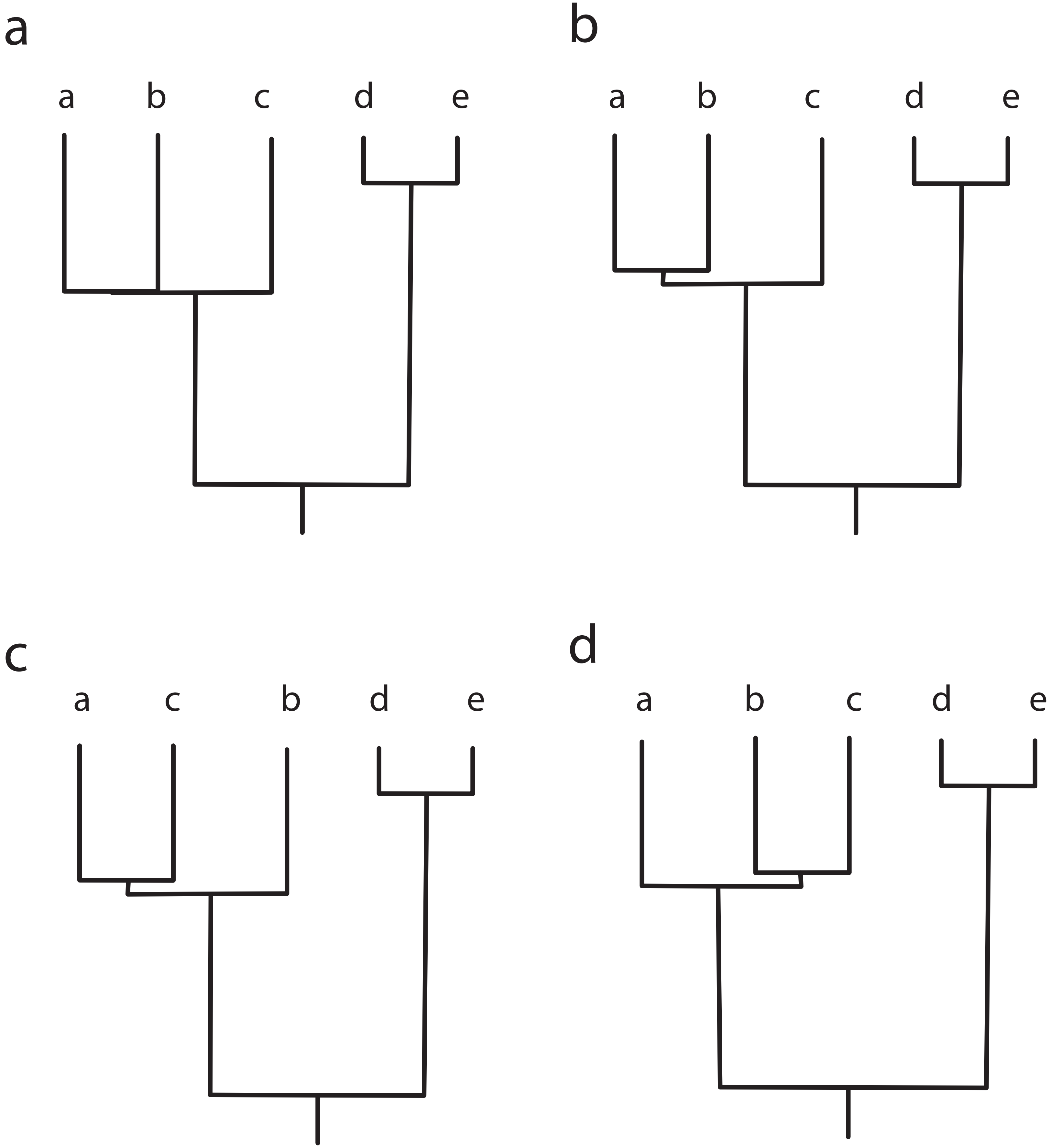
PHYLOGENETIC TREES



PHYLOGENETIC TREES



PHYLOGENETIC TREES



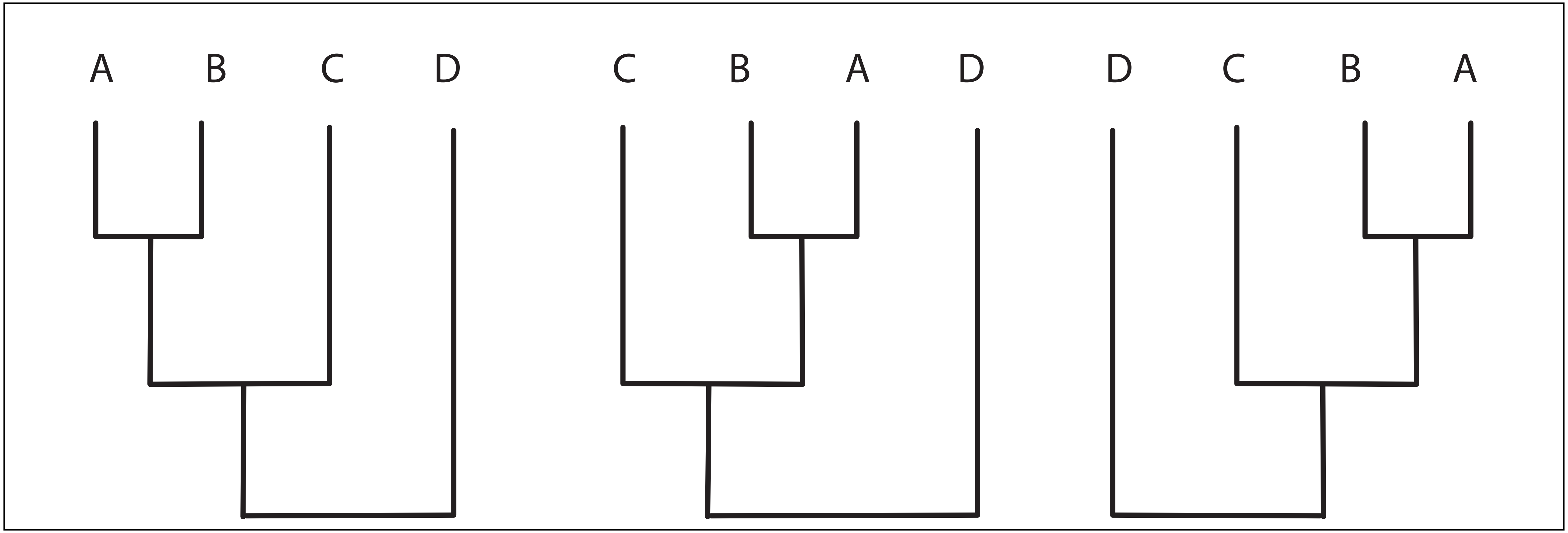
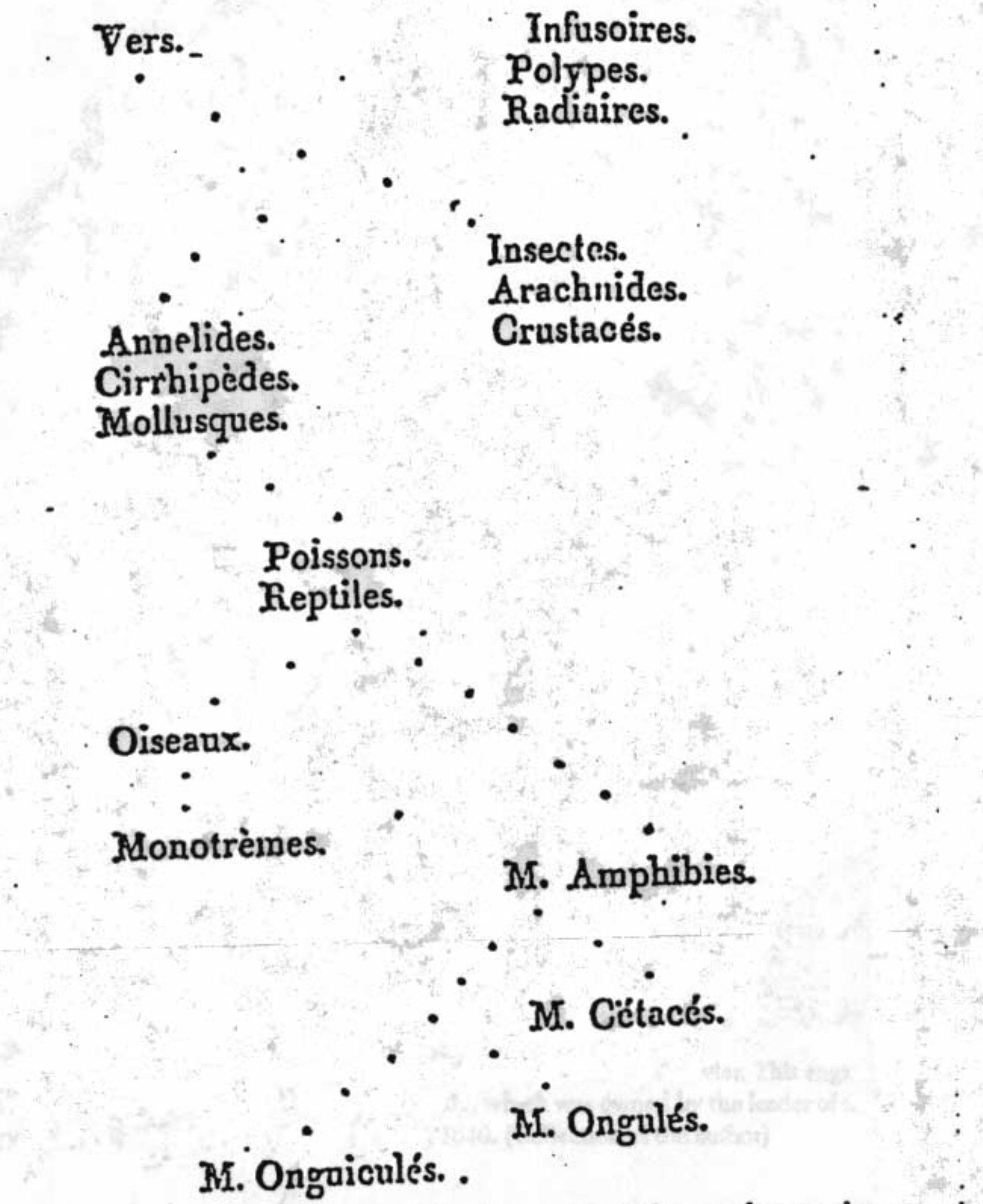


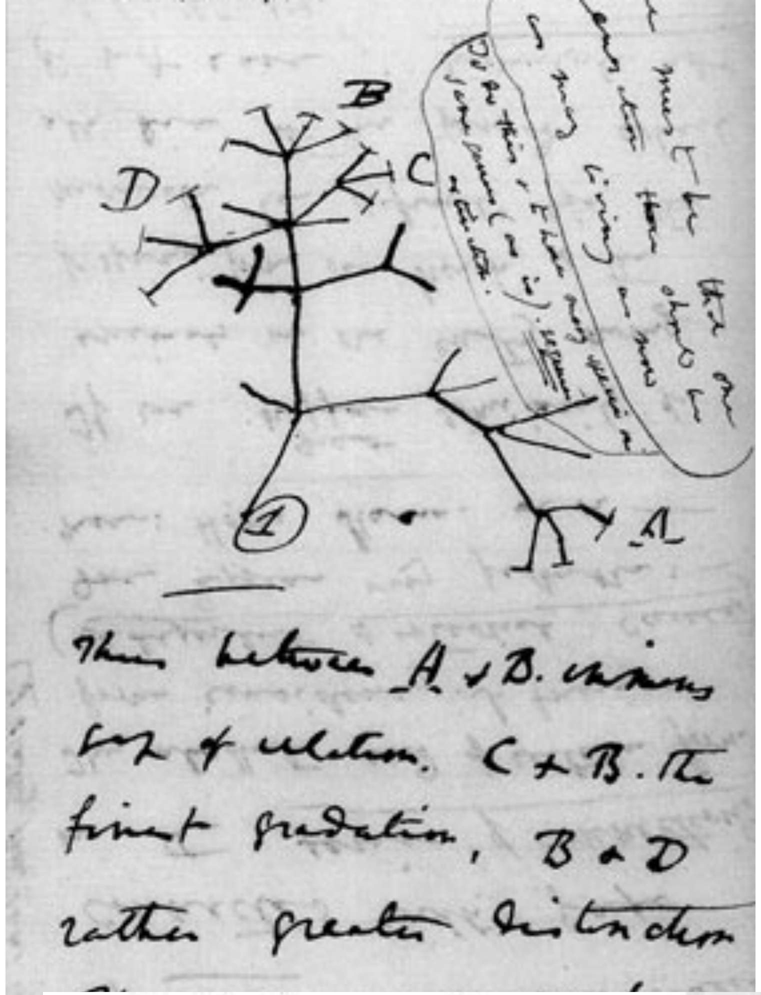
Figure 7: The same tree can be drawn in many different ways because of the rotational freedom around each node.

TABLEAU

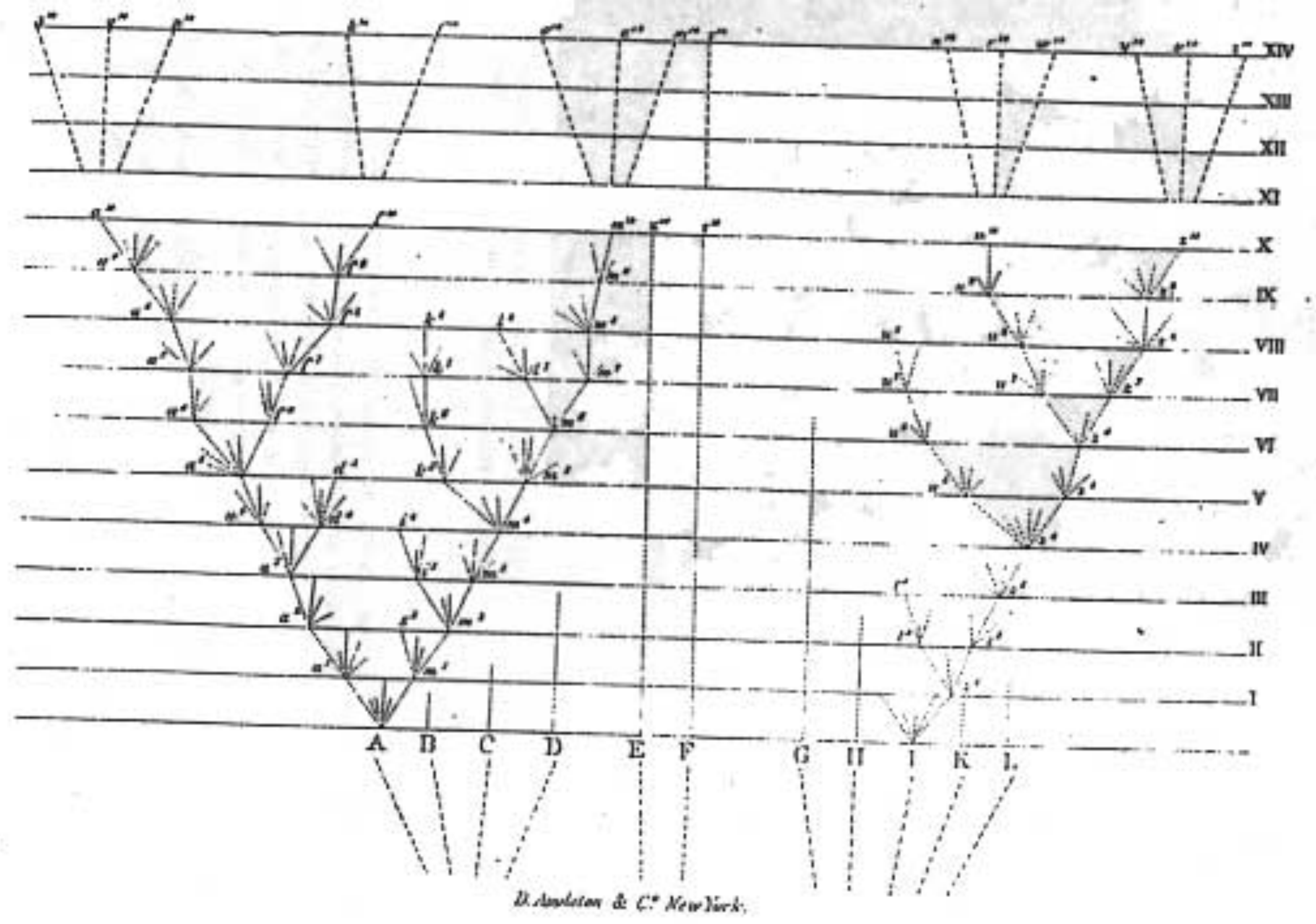
Servant à montrer l'origine des différens animaux.



The first evolutionary tree, upside down from the modern point of view, published in Lamarck's *Philosophic 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."

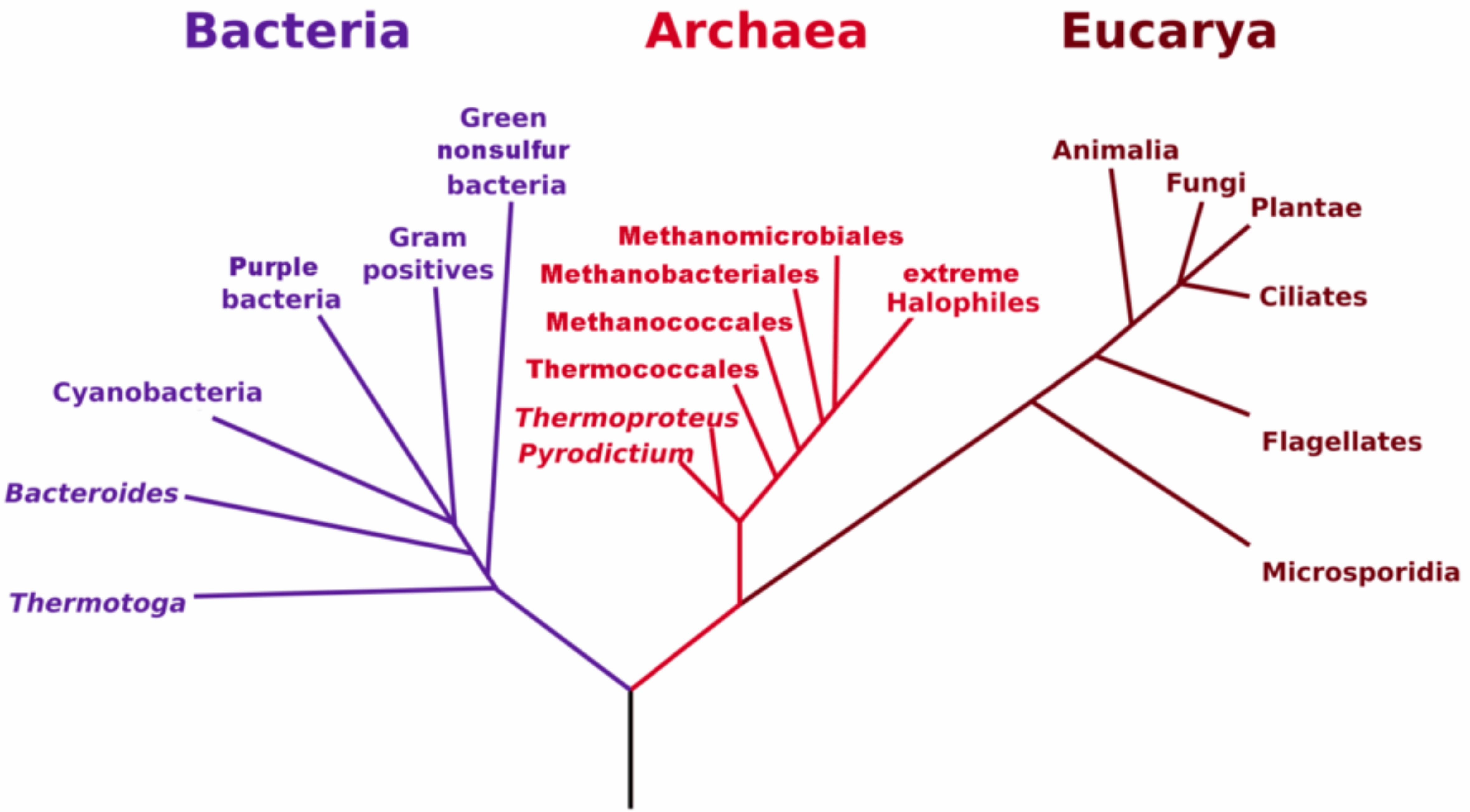

must be that one
generation then should be
as many living as now
To do this & to have many species in
same genus (as is). requires extinction.
B
C
D
1
A

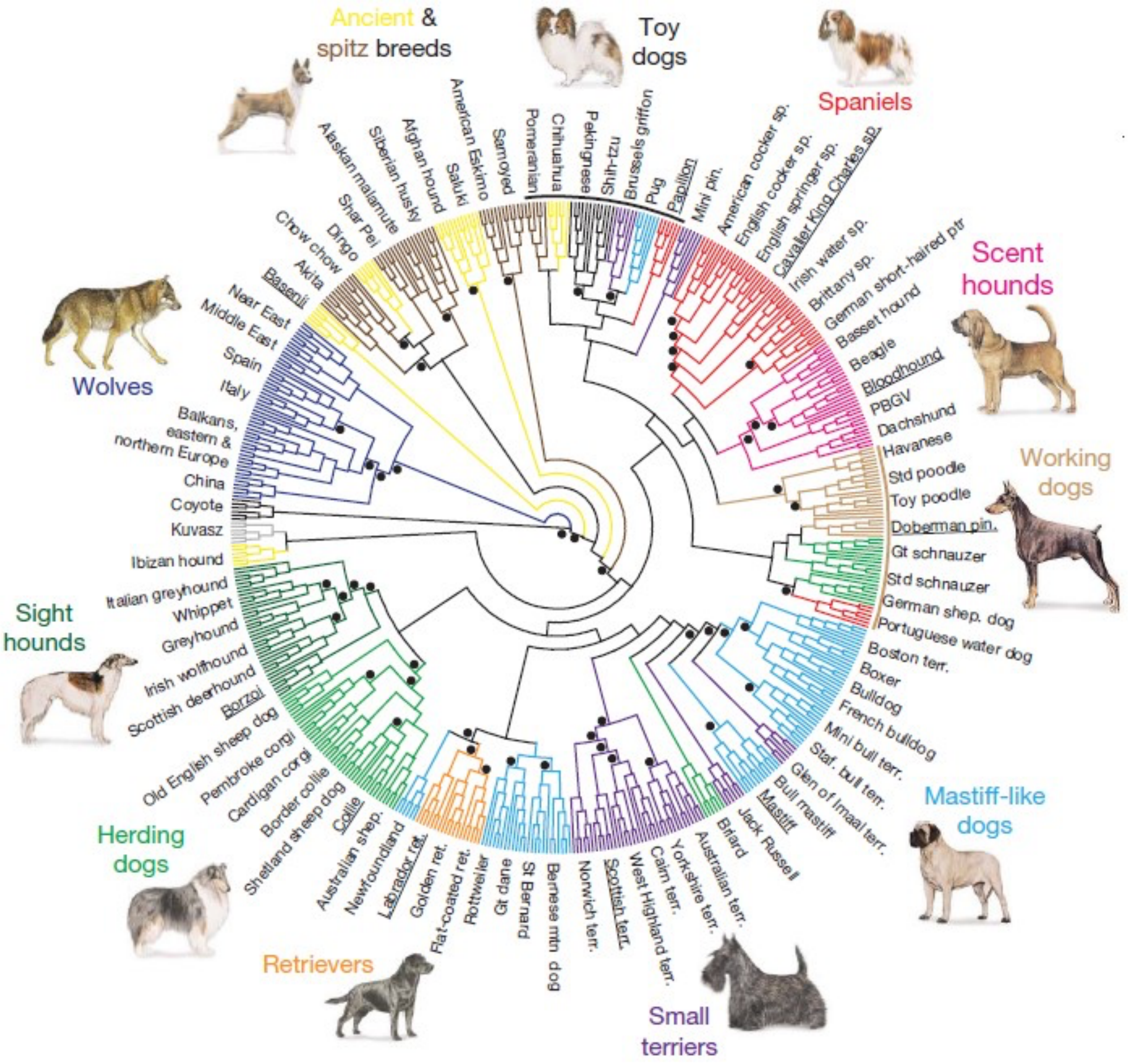
Thus between A & B. immens gap of relation. C & B. the finest gradation, B & D rather greater distinction

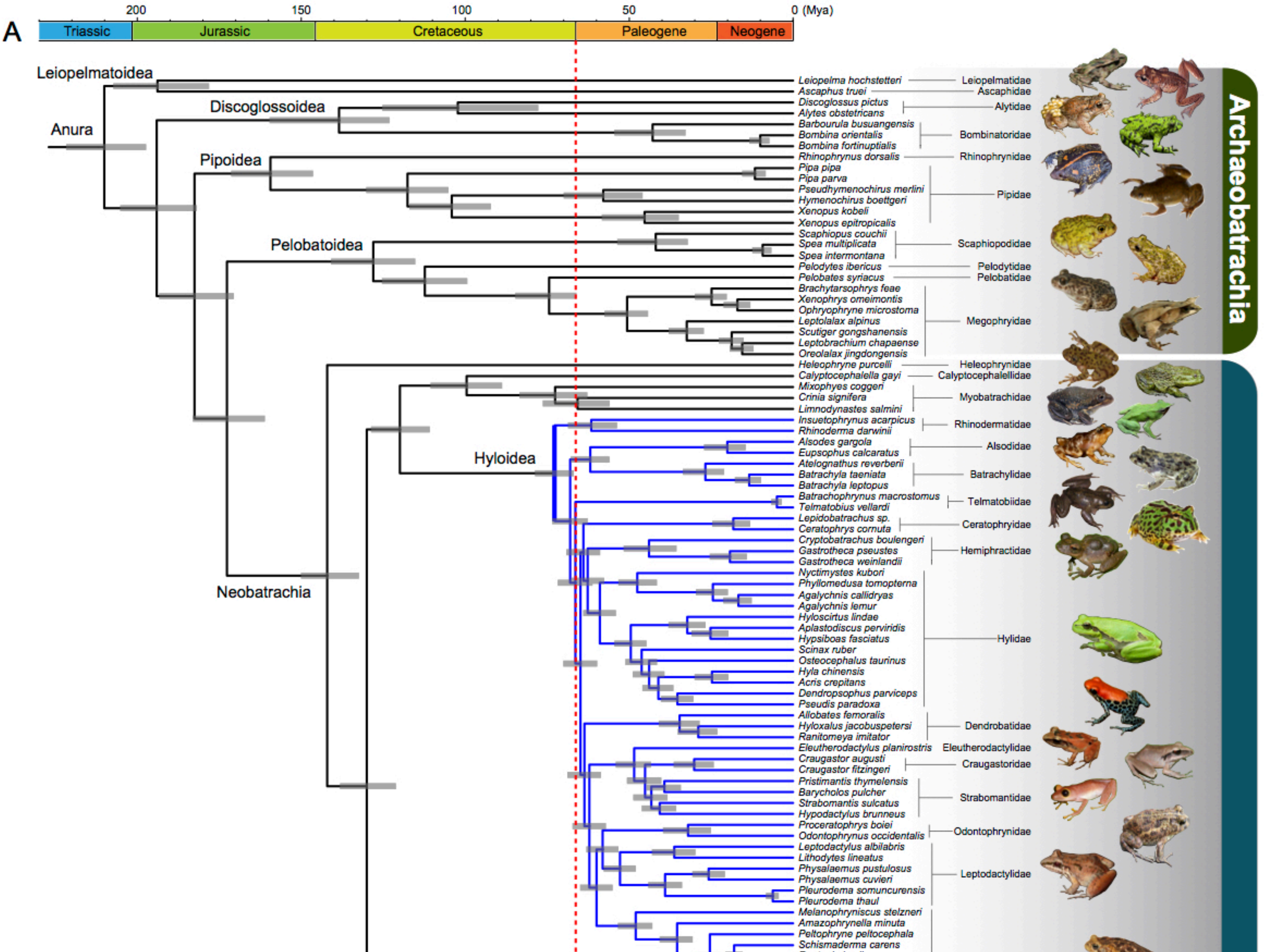


D. Appleton & Co. New York.

Phylogenetic Tree of Life

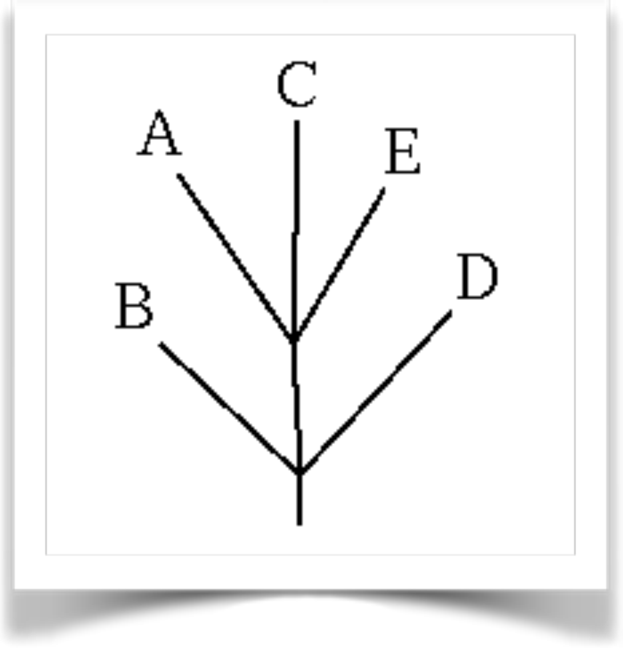



Ancient &
spitz breeds
Toy
dogs
Spaniels
Scent
hounds
Working
dogs
Mastiff-like
dogs
Small
terriers
Retrievers
Herding
dogs
Sight
hounds
Wolves
Chow chow
Dingo
Shar Pei
Alaskan malamute
Siberian husky
Afghan hound
Saluki
American Eskimo
Samoyed
Pomeranian
Chihuahua
Pekingese
Shih-tzu
Brussels griffon
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.
Scottish terr.
Norwich 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



A
200
150
100
50
0 (Mya)
Triassic
Jurassic
Cretaceous
Paleogene
Neogene
Leiopelmatoidea
Anura
Discoglossoidea
Pipoidea
Pelobatoidea
Neobatrachia
Hyloidea
Leiopelma hochstetteri
Ascaphus truei
Discoglossus pictus
Alytes obstetricans
Barbourula busuangensis
Bombina orientalis
Bombina fortinuptialis
Rhinophrynus dorsalis
Pipa pipa
Pipa parva
Pseudhymenochirus merlini
Hymenochirus boettgeri
Xenopus kobeli
Xenopus epitropicalis
Scaphiopus couchii
Spea multiplicata
Spea intermontana
Pelodytes ibericus
Pelobates syriacus
Brachytarsophrys feae
Xenophrys omeimontis
Ophryophryne microstoma
Leptolalax alpinus
Scutiger gongshanensis
Leptobrachium chapaense
Oreolalax jingdongensis
Heleophryne purcelli
Calyptocephalella gayi
Mixophyes coggeri
Crinia signifera
Limnodynastes salmini
Insuetophrynus acarpicus
Rhinoderma darwinii
Alsodes gargola
Eupsophus calcaratus
Atelognathus reverberii
Batrachyla taeniata
Batrachyla leptopus
Batrachophrynus macrostomus
Telmatobius vellardi
Lepidobatrachus sp.
Ceratophrys cornuta
Cryptobatrachus boulengeri
Gastrotheca pseustes
Gastrotheca weinlandii
Nyctimystes kubori
Phyllomedusa tomopterna
Agalychnis callidryas
Agalychnis lemur
Hyloscirtus lindae
Aplastodiscus perviridis
Hypsiboas fasciatus
Scinax ruber
Osteocephalus taurinus
Hyla chinensis
Acris crepitans
Dendropsophus parviceps
Pseudis paradoxa
Allobates femoralis
Hyloxalus jacobuspetersi
Ranitomeya imitator
Eleutherodactylus planirostris
Craugastor augusti
Craugastor fitzingeri
Pristimantis thymelensis
Barycholos pulcher
Strabomantis sulcatus
Hypodactylus brunneus
Proceratophrys boiei
Odontophrynus occidentalis
Leptodactylus albilabris
Lithodytes lineatus
Physalaemus pustulosus
Physalaemus cuvieri
Pleurodema somuncurensis
Pleurodema thaul
Melanophryniscus stelzneri
Amazophrynella minuta
Peltophryne peltocephala
Schismaderma carens
Leiopelmatidae
Ascaphidae
Alytidae
Bombinatoridae
Rhinophrynidae
Pipidae
Scaphiopodidae
Pelodytidae
Pelobatidae
Megophryidae
Heleophrynidae
Calyptocephalellidae
Myobatrachidae
Rhinodermatidae
Alsodidae
Batrachylidae
Telmatobiidae
Ceratophryidae
Hemiphractidae
Hylidae
Dendrobatidae
Eleutherodactylidae
Craugastoridae
Strabomantidae
Odontophrynidae
Leptodactylidae
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,,,);

NEWICK TREE FORMAT

(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);
```

NEWICK TREE FORMAT

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

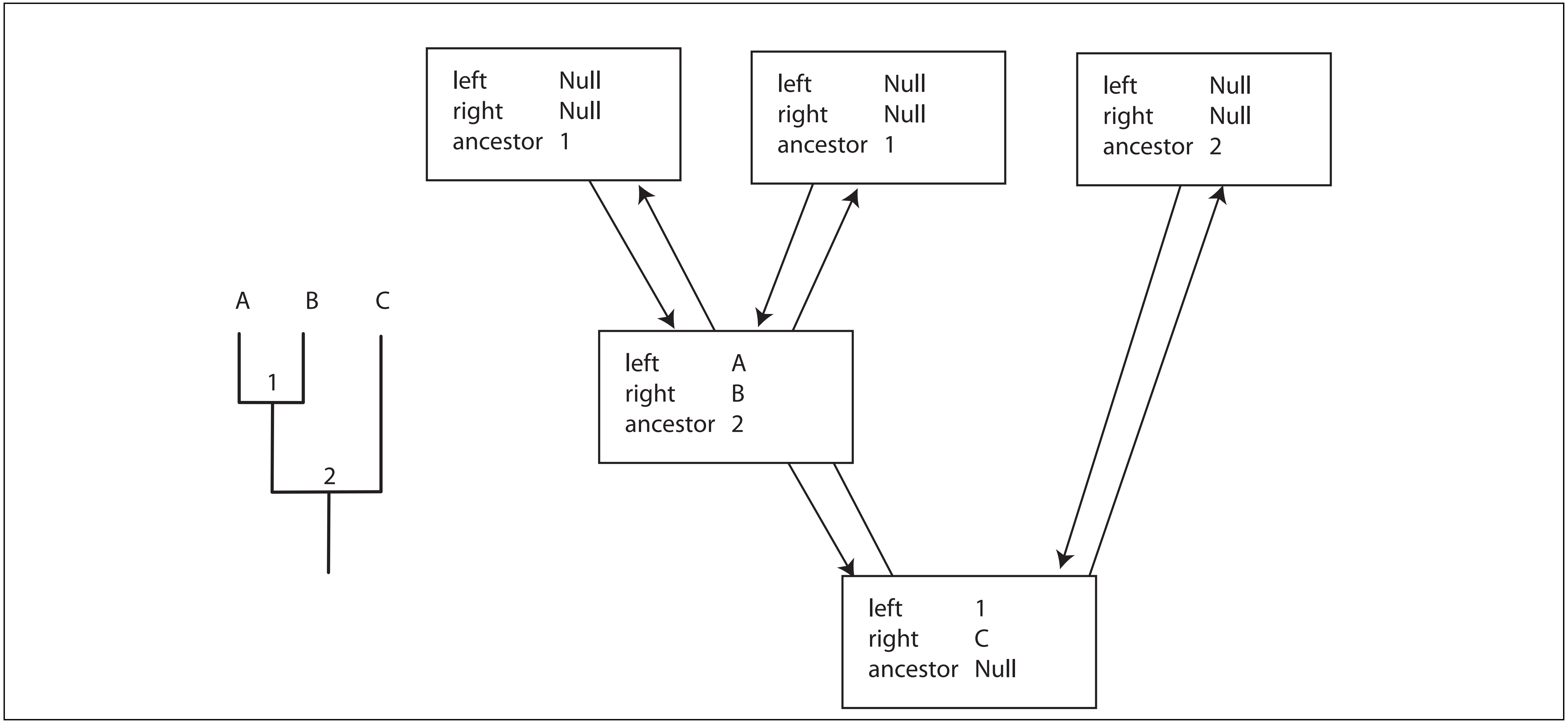
```
(A,(B,C),D);
```

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.



left Null
right Null
ancestor 1

left Null
right Null
ancestor 1

left Null
right Null
ancestor 2

left A
right B
ancestor 2

left 1
right C
ancestor Null

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

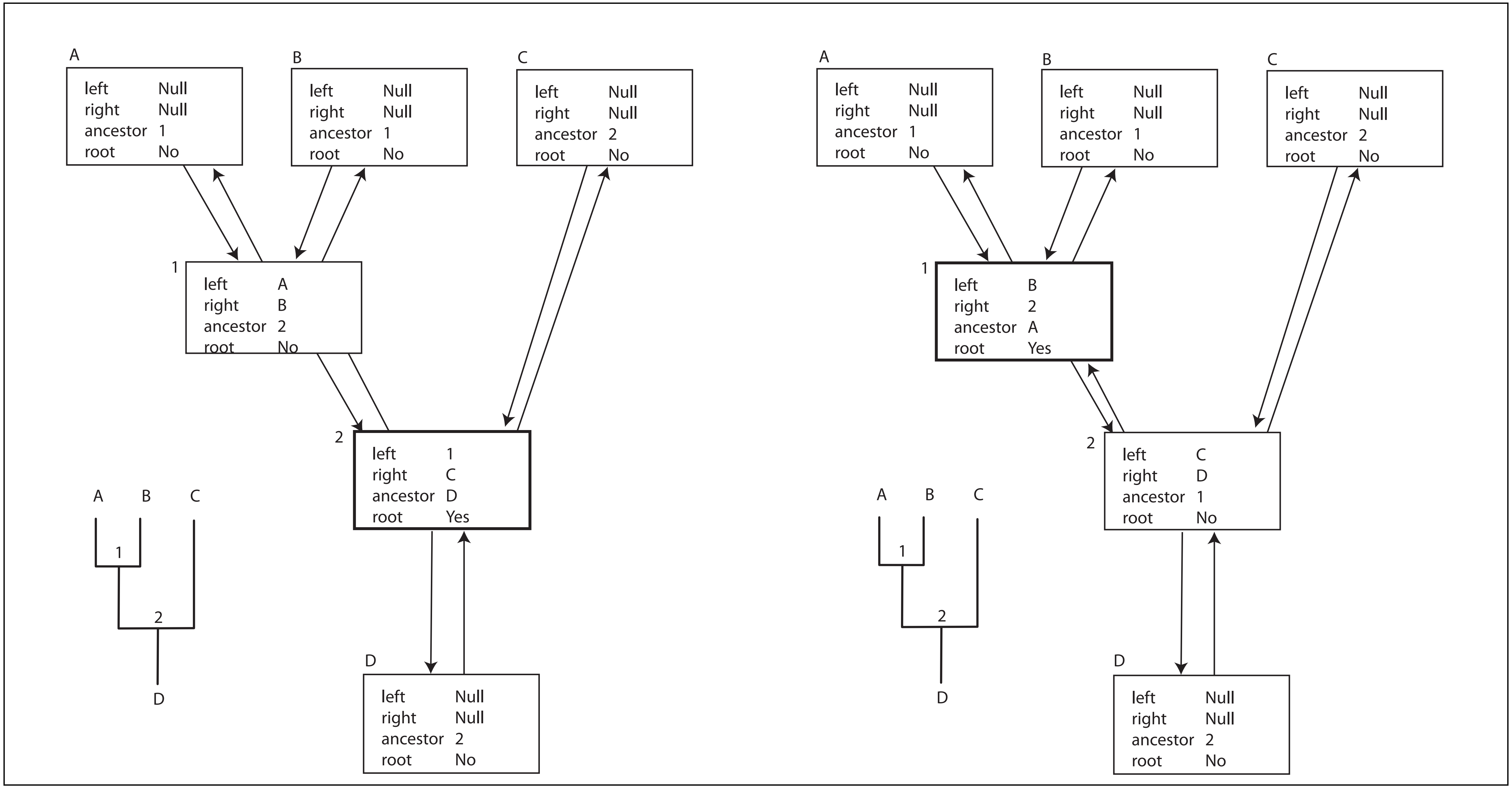


Figure 9: A modification of BTM to accommodate unrooted trees. The two trees have different roots, the pointers change in all interior nodes.

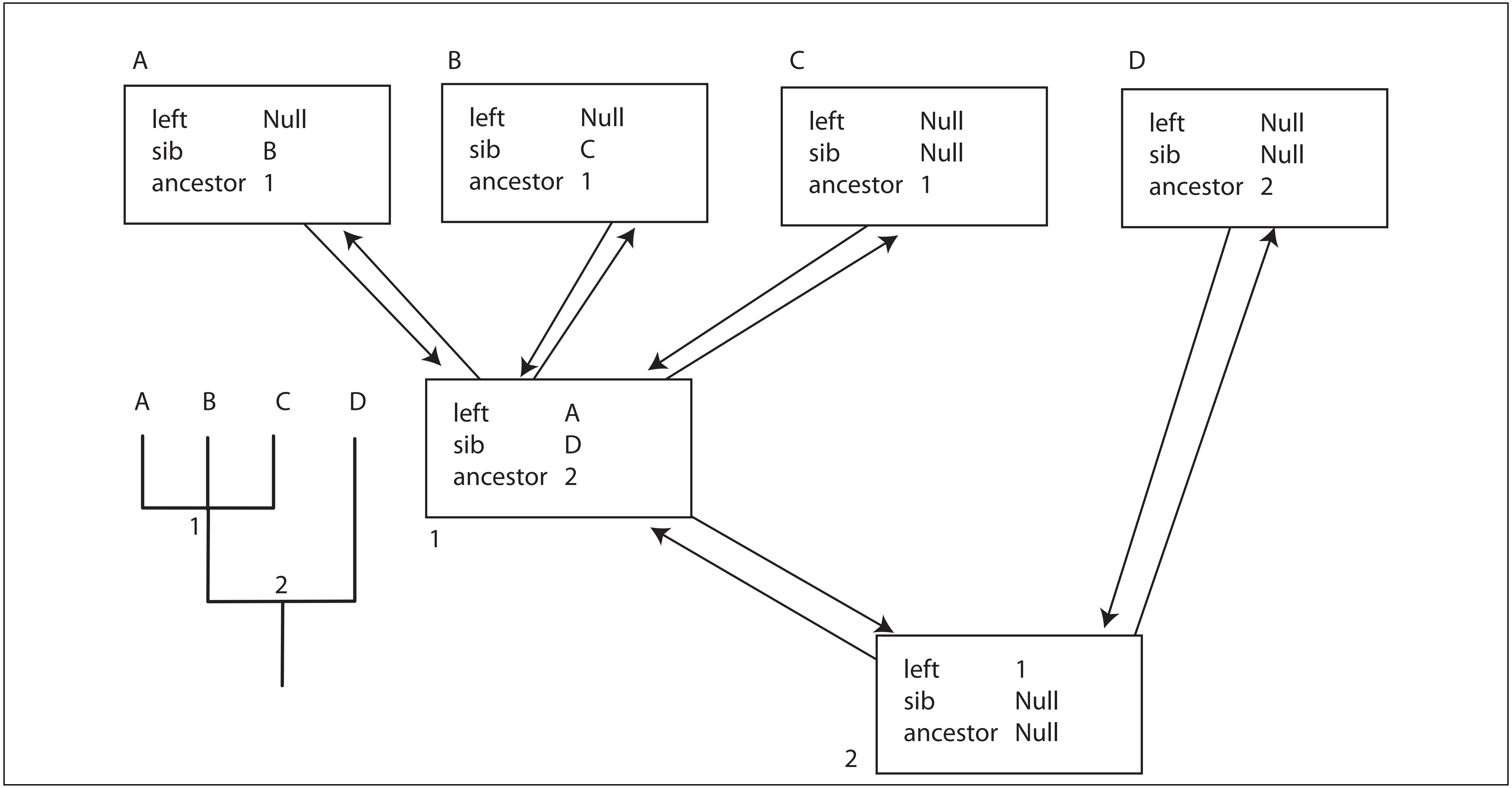


Figure 10: The polytomous tree data model (PTM).

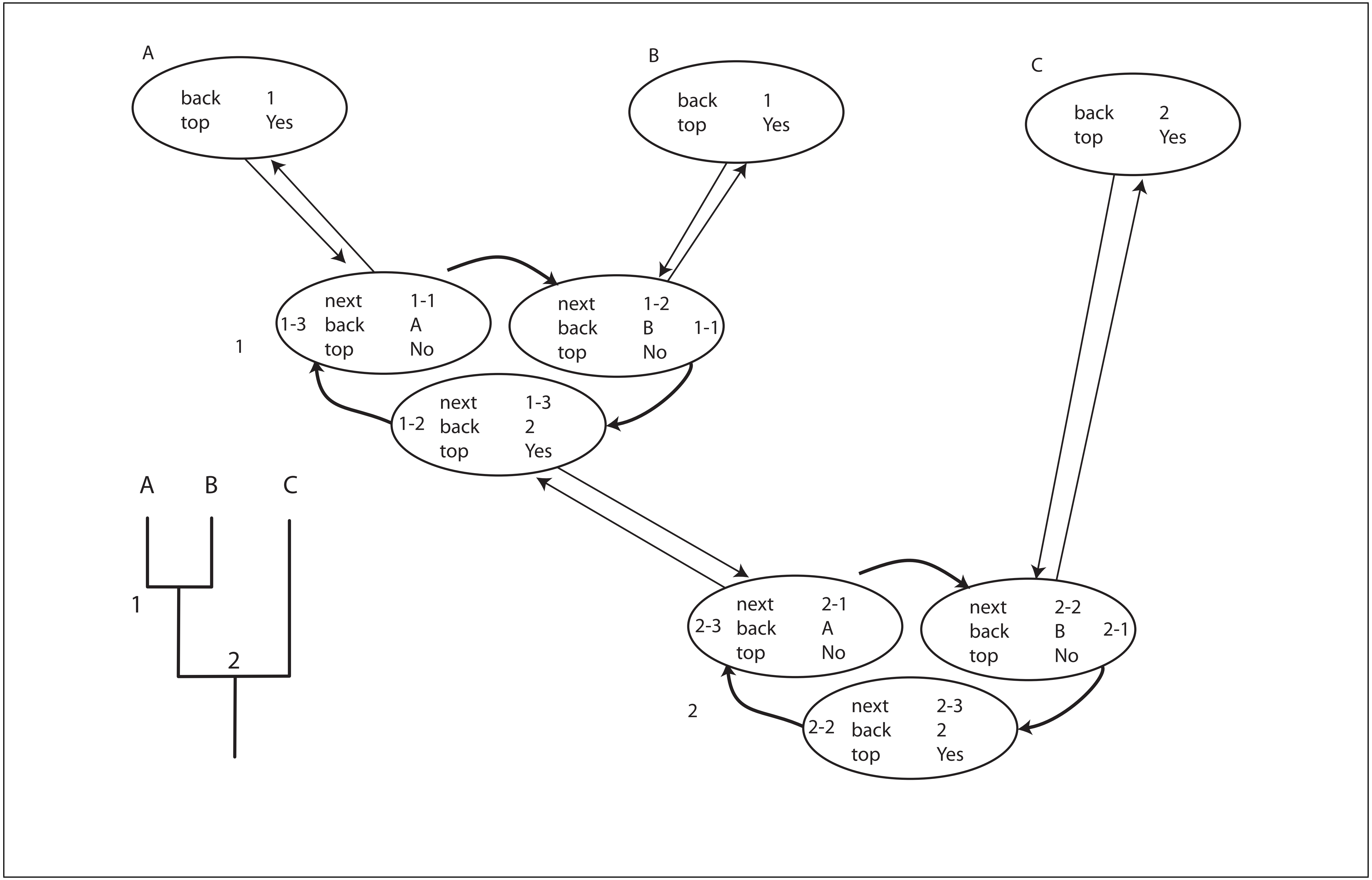


Figure 11: The Felsenstein tree data model (FTM). Each node is broken up into nodelets, this facilitates greatly rerooting and polytomous trees.

PHYLOGENETIC 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 $p$
```

PHYLOGENETIC TREES

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)
```