BIOINFORMATICS AND STATISTICAL GENETICS

GABRIEL VALIENTE

ALGORITHMS, BIOINFORMATICS, COMPLEXITY AND FORMAL METHODS RESEARCH GROUP,
TECHNICAL UNIVERSITY OF CATALONIA

2023–2024

Phylogenetic reconstruction II

- Phylogenies and taxonomies
- Classification of metagenomic samples
- The taxonomic assignment problem
- Accuracy and coverage
- The LCA skeleton tree

- K
- P
- C
- O
- F
- G
- S

- King
- Philip
- Came
- Over
- For
- Good
- Soup

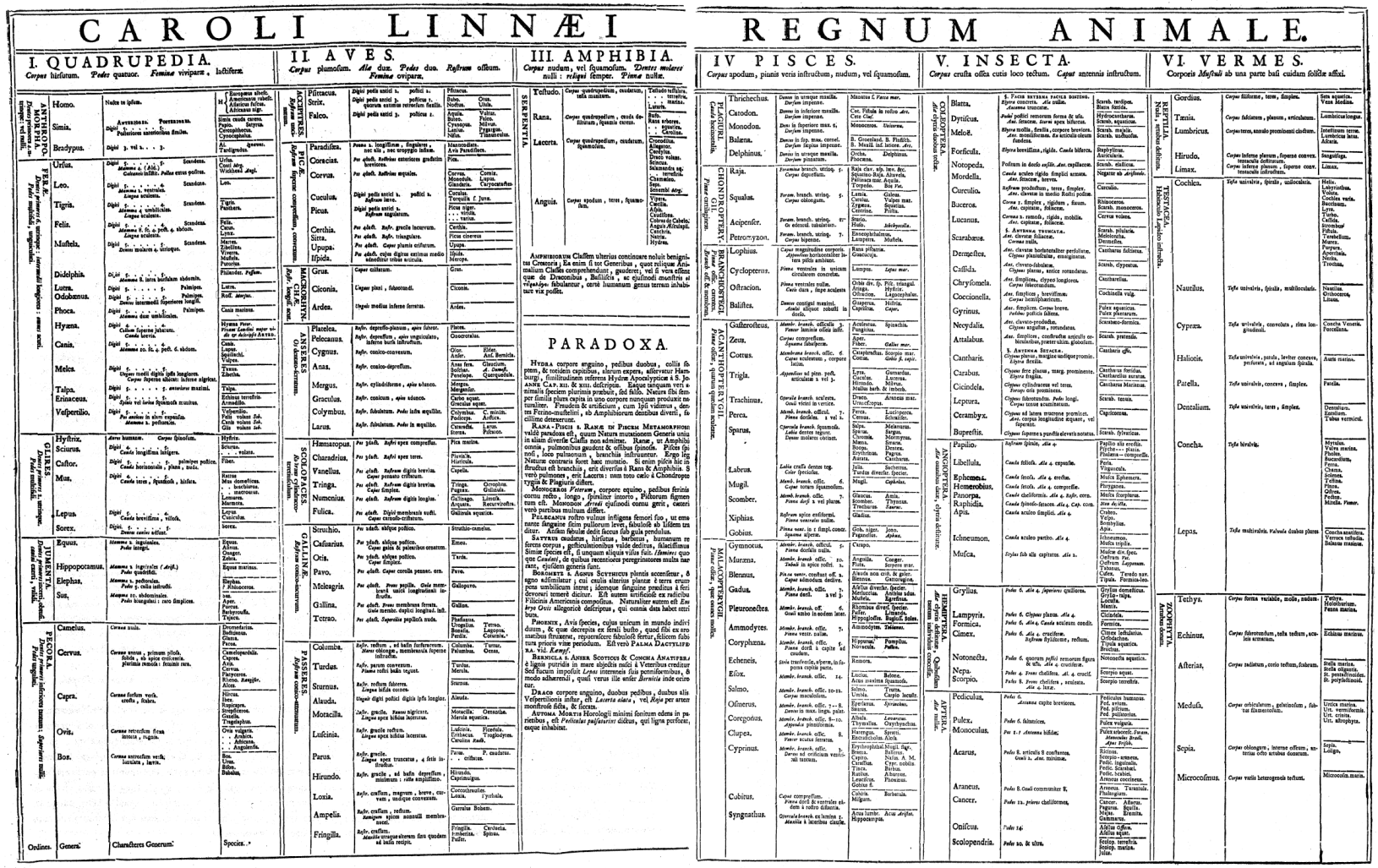
- Kingdom
- Phylum
- Class
- Order
- Family
- Genus
- Species

- The classifying of living things is called taxonomy
- Taxonomic rank shows how a species fits into bigger groups

- One of the reasons a land tortoise is classed as a reptile (class Reptilia) is that it hatches from a soft-shelled egg
- Most reptiles have this kind of egg
- Biologists decide which large group a species belongs to by looking for shared characteristics like this one

- There are now twice as many species of lemur than there used to be
- This is not because we have discovered many more kinds of lemur but because we discovered DNA
- Some biologists now consider a species to be distinct from a similar one if just 2 per cent of its DNA is unique
- Biologists now recognize more than 100 species of lemur

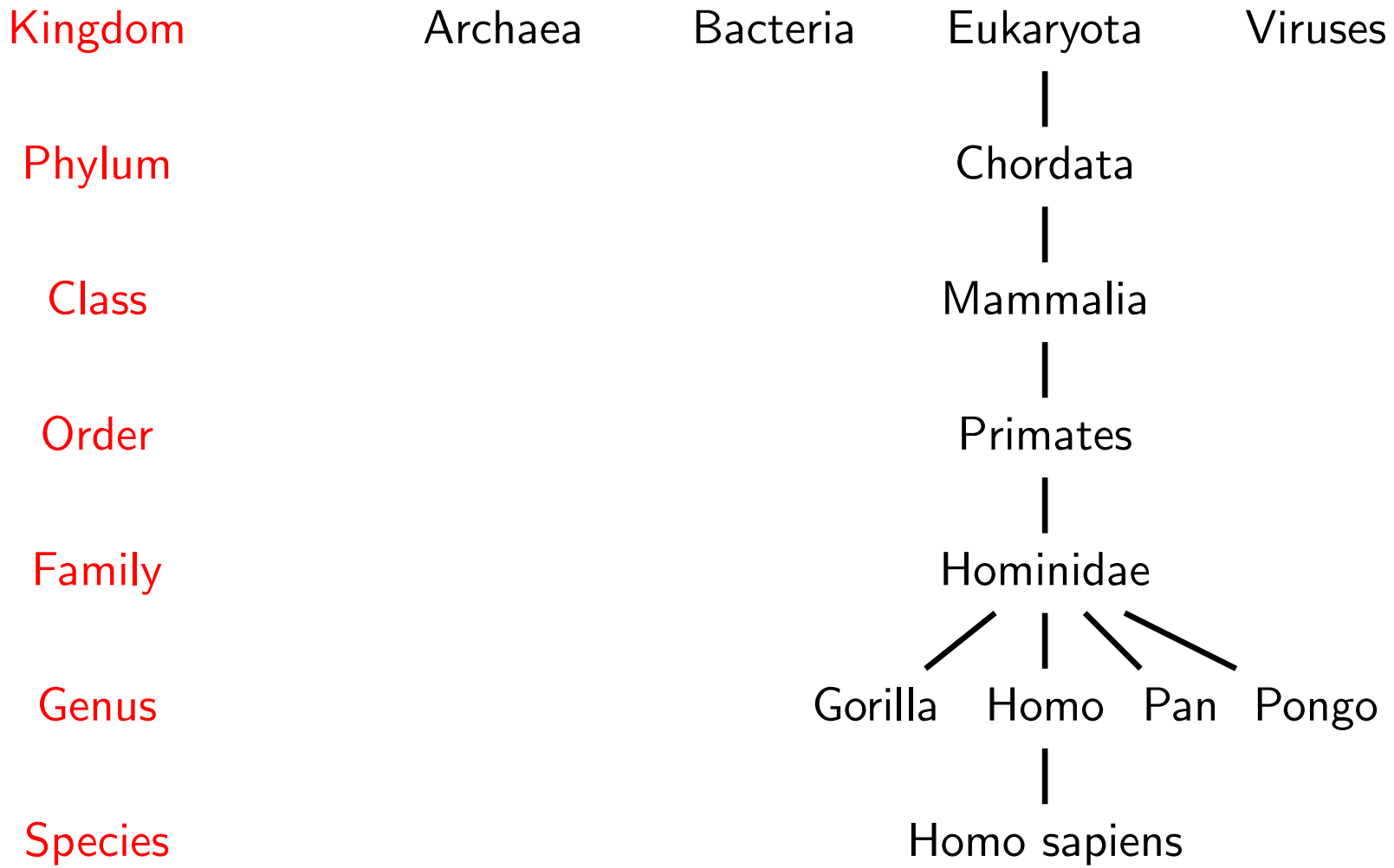
- C. Lloyd, editor. *Britannica All New Children's Encyclopedia*. Britannica Books, Kent, UK, 2020

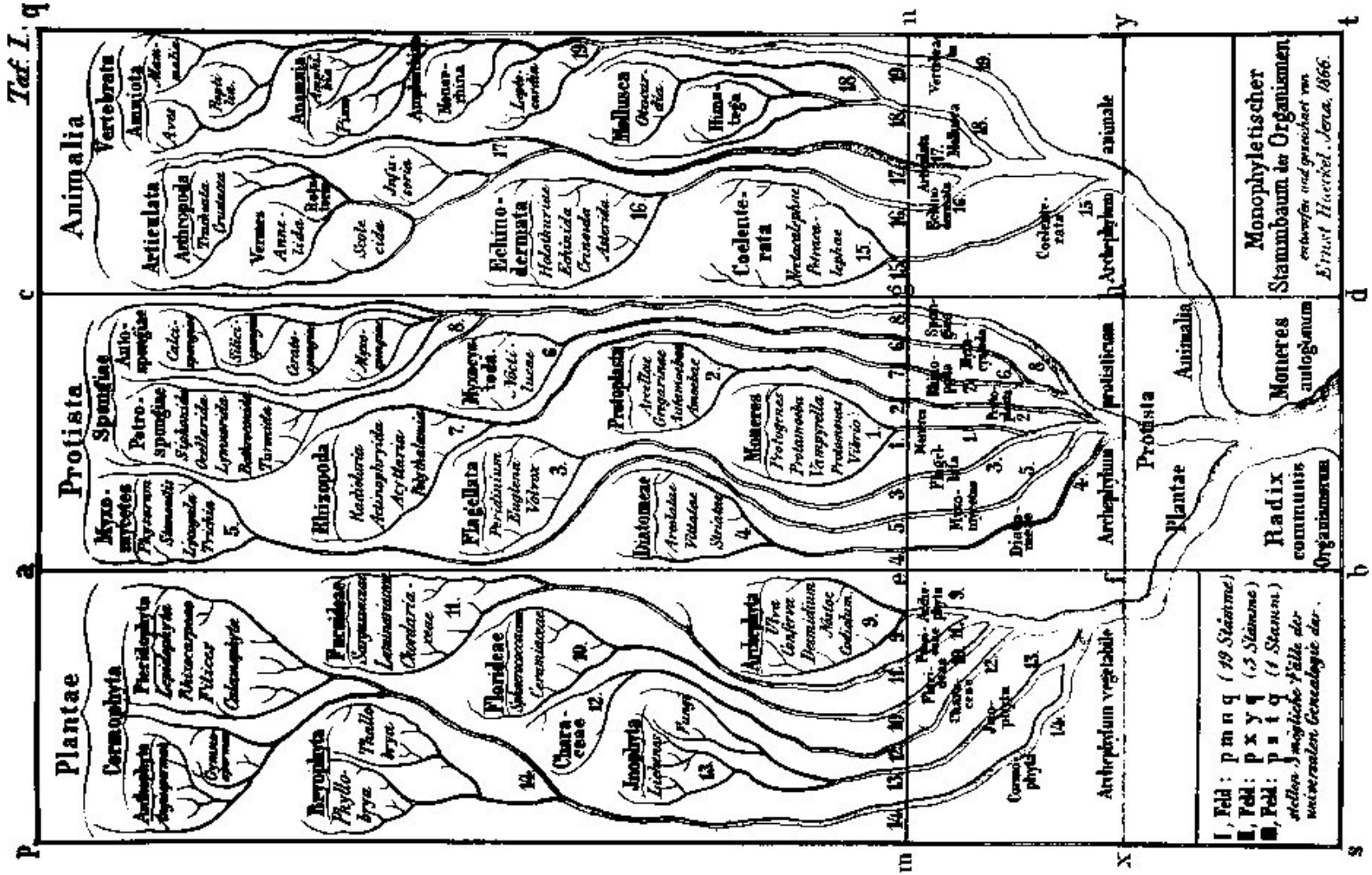


- C. Linnæus. *Systema Naturae*. L. Salvius, Stockholm, 1735

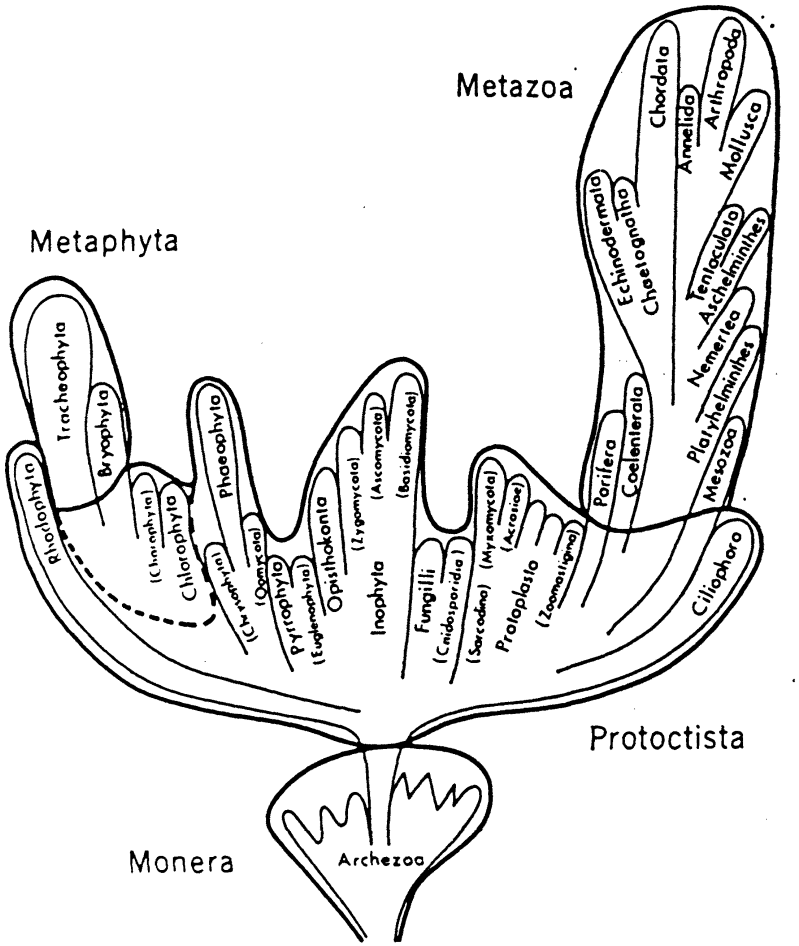
Kingdom	Archaea	Bacteria	Eukaryota	Viruses
Phylum			Streptophyta	
Class			Streptophytina	
Order			Solanales	
Family			Solanaceae	
Genus			Solanum	
Species			Solanum lycopersicum	

Solanum caule inermi herbaceo, foliis pinnatis incisis, racemis simplicibus

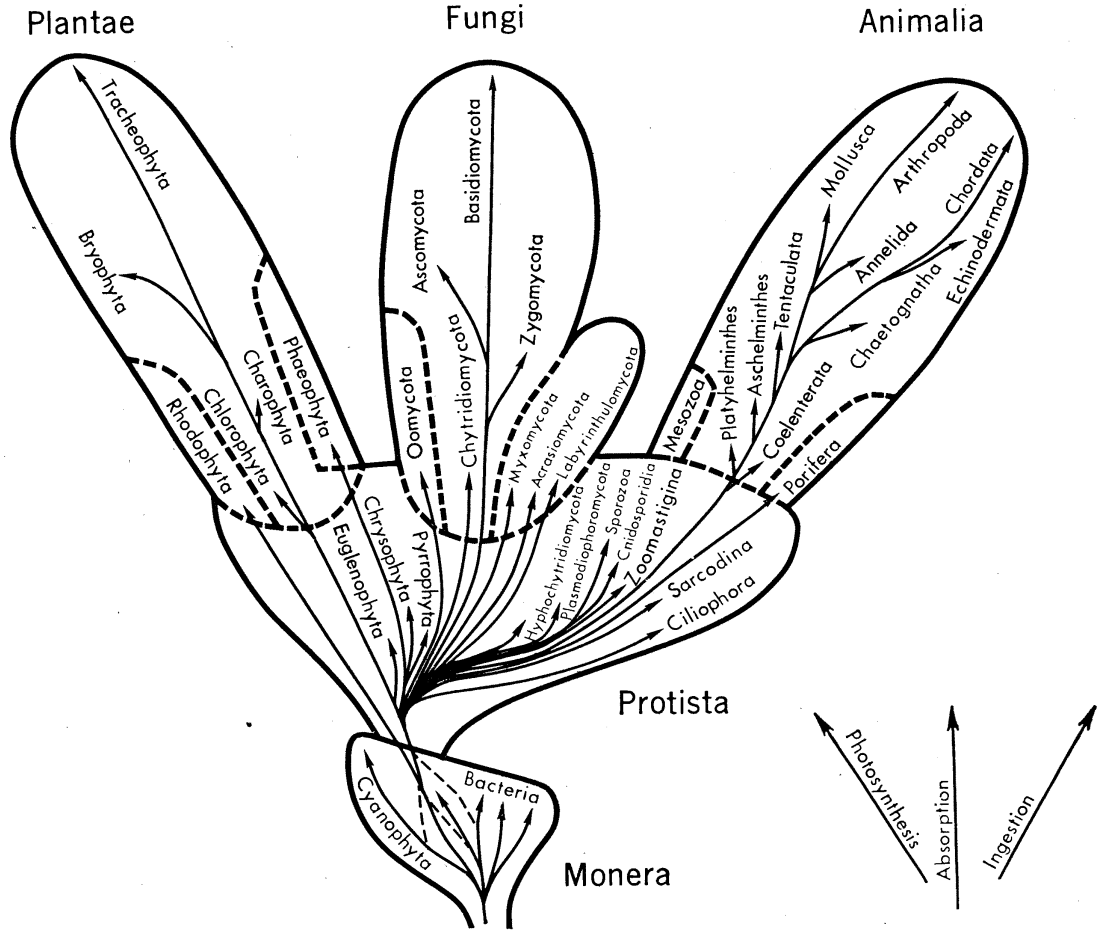

Kingdom
Archaea
Bacteria
Eukaryota
Viruses
Phylum
Chordata
Class
Mammalia
Order
Primates
Family
Hominidae
Genus
Gorilla
Homo
Pan
Pongo
Species
Homo sapiens



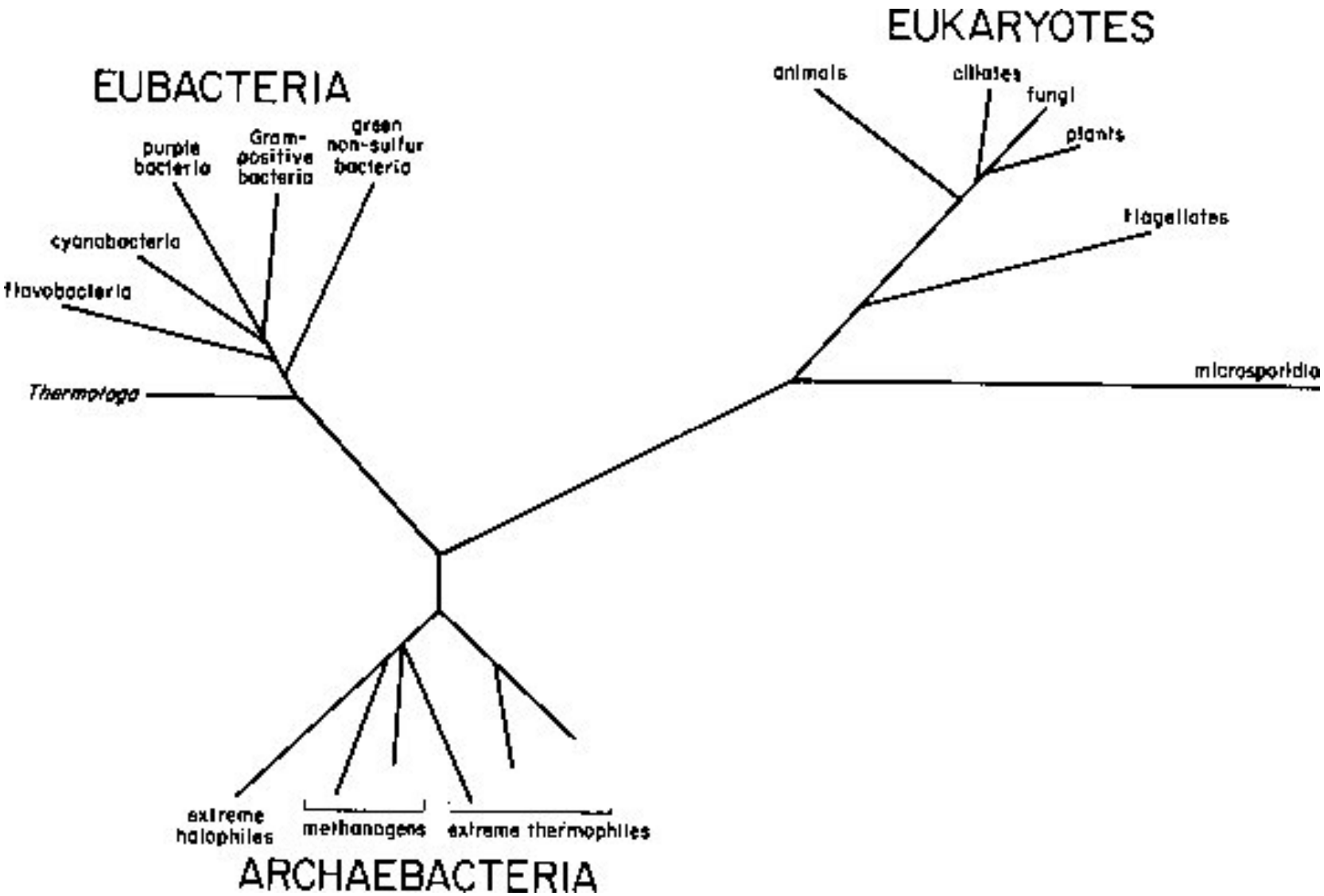
- E. Haeckel. *Generelle Morphologie der Organismen*. Georg Reimer, Berlin, 1866



- H. F. Copeland. The kingdoms of organisms. *The Quarterly Review of Biology*, 13(4):383–420, 1938



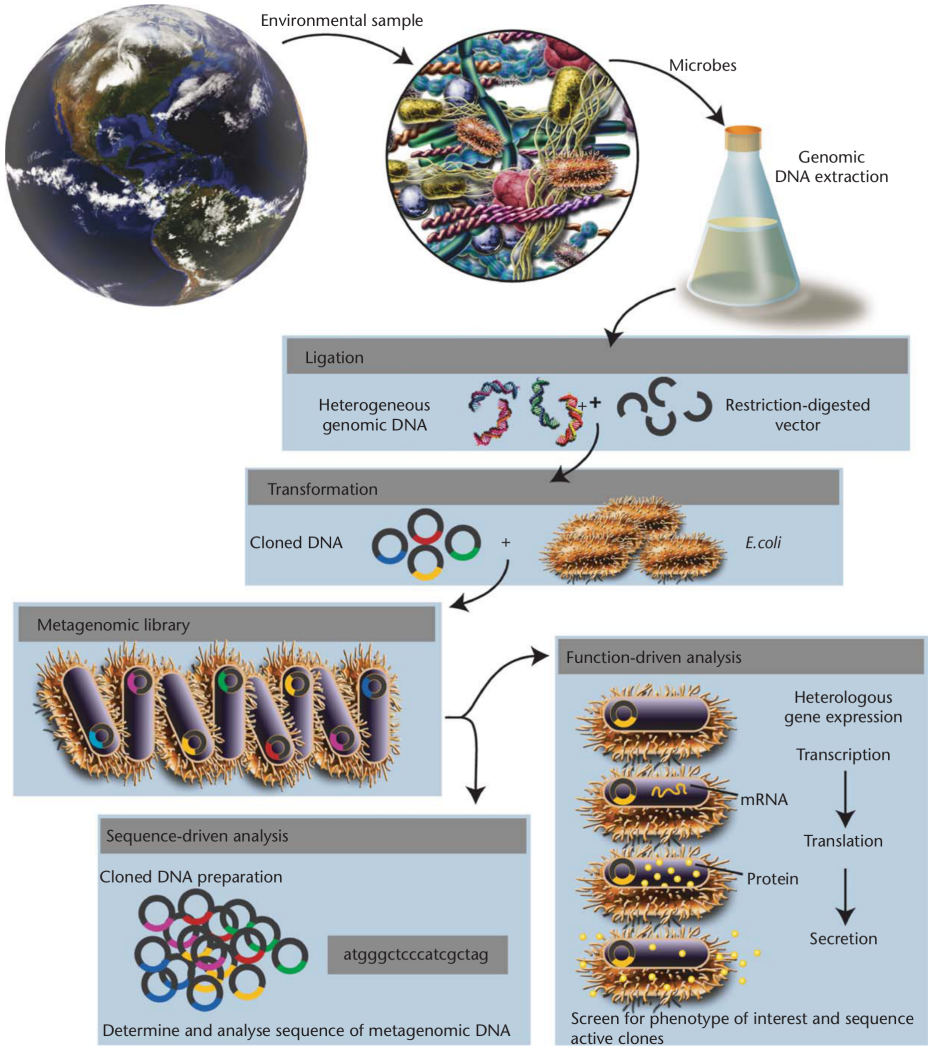
- R. H. Whittaker. New concepts of kingdoms of organisms. *Science*, 163(3863):150–160, 1969



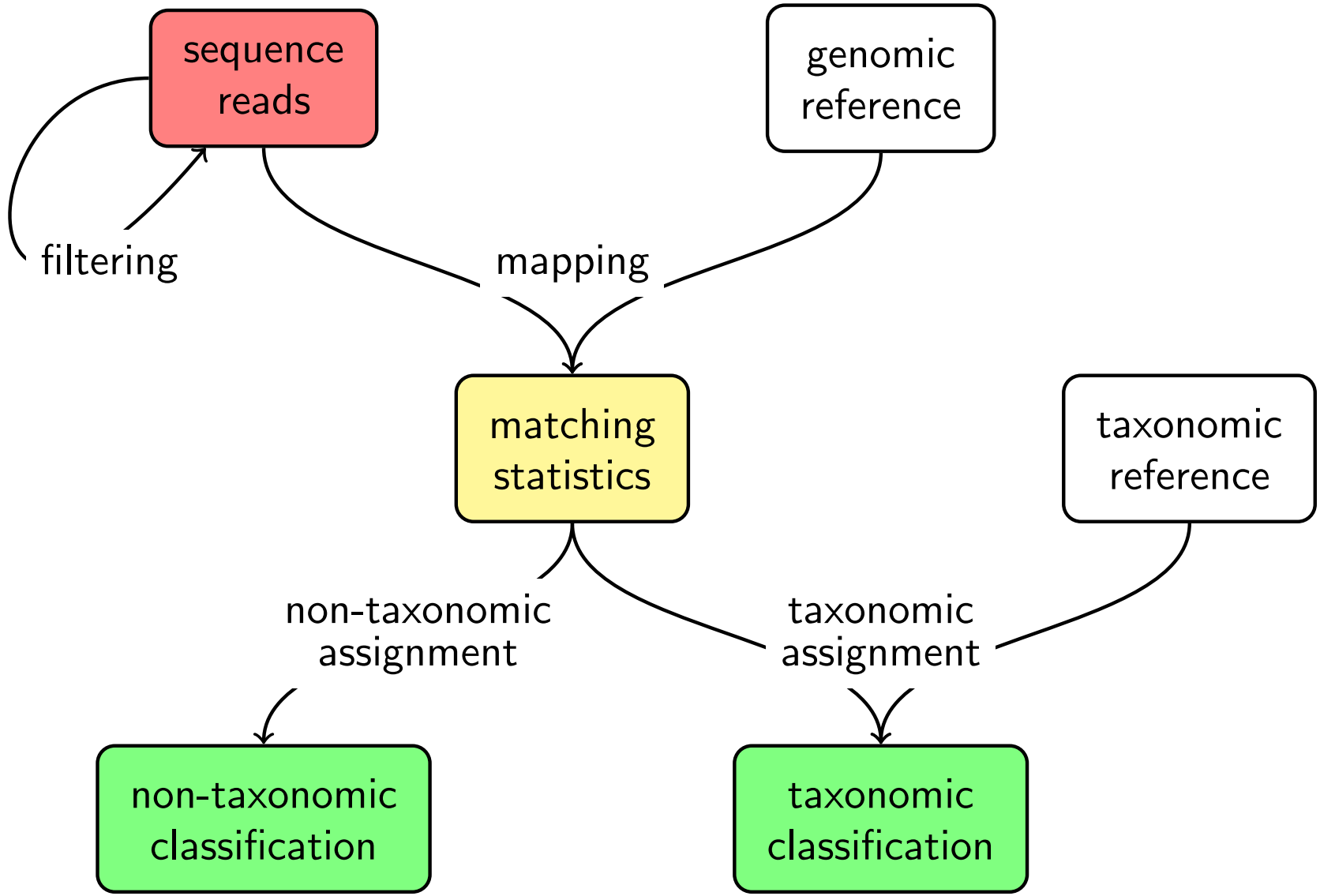
- C. R. Woese. Bacterial evolution. *Microbiology Reviews*, 51(2):221–271, 1987

<table>
<tr><th>Linnæus
1735</th><th>Haeckel
1866</th><th>Copeland
1938</th><th>Whittaker
1969</th><th>Woese
1987</th></tr>
<tr><td rowspan="3"></td><td rowspan="3">Protista</td><td rowspan="2">Monera</td><td rowspan="2">Monera</td><td>Bacteria</td></tr>
<tr><td>Archaea</td></tr>
<tr><td>Protoctista</td><td>Protista</td><td rowspan="4">Eukarya</td></tr>
<tr><td rowspan="2">Plantae</td><td rowspan="2">Plantae</td><td>Plantae</td><td>Plantae</td></tr>
<tr><td>Protoctista</td><td>Fungi</td></tr>
<tr><td>Animalia</td><td>Animalia</td><td>Animalia</td><td>Animalia</td></tr>
</table>

- S. Federhen. The NCBI taxonomy database. *Nucleic Acids Research*, 40(D1):D136–D143, 2012
- J. R. Cole et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, 41(D1):D633–D642, 2014
- D. McDonald et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, 6(3):610–618, 2012



- J. Handelsman. Metagenomics and microbial communities. In *Encyclopedia of Life Sciences*. John Wiley & Sons, 2007


sequence reads
genomic reference
filtering
mapping
matching statistics
taxonomic reference
non-taxonomic assignment
taxonomic assignment
non-taxonomic classification
taxonomic classification

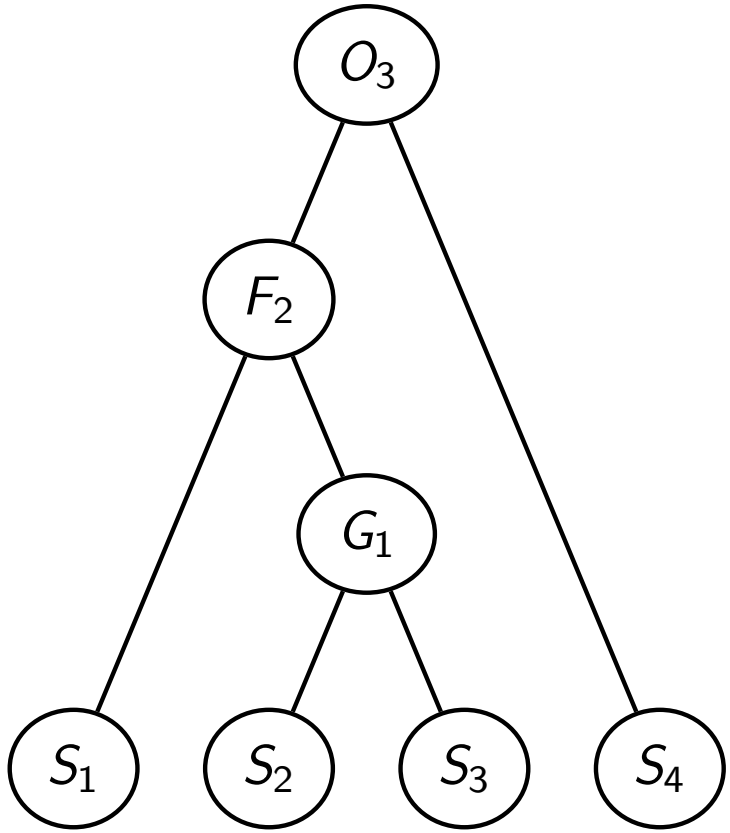
Classification of reads from a metagenomic sample using a reference taxonomy

- Mapping the reads to the reference sequences
- Classifying each read at a node under the LCA of the candidate sequences in the reference taxonomy with the least classification error

Potential sources of bias

- Imbalanced reference taxonomy
- Multiple nodes in the reference taxonomy with the least classification error for a given read

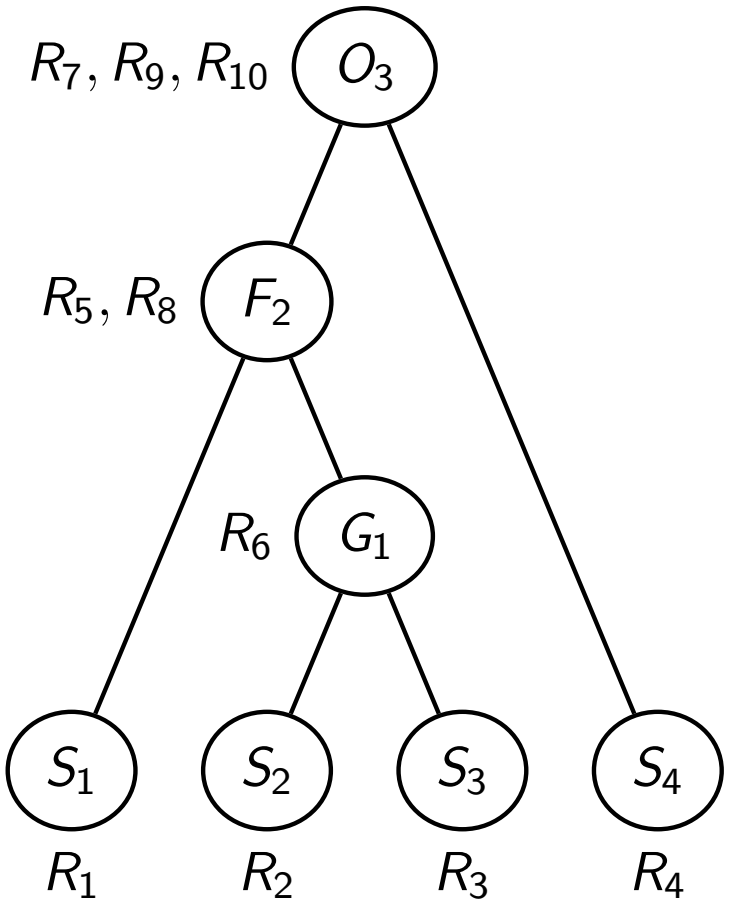
	S_1	S_2	S_3	S_4
R_1	x			
R_2		x		
R_3			x	
R_4				x
R_5	x	x		
R_6		x	x	
R_7			x	x
R_8	x	x	x	
R_9		x	x	x
R_{10}	x	x	x	x



- Reads $R_1, \ldots, R_{10}$ match sequences $S_1, \ldots, S_4$

LCA mapping

	S_1	S_2	S_3	S_4
R_1	x			
R_2		x		
R_3			x	
R_4				x
R_5	x	x		
R_6		x	x	
R_7			x	x
R_8	x	x	x	
R_9		x	x	x
R_{10}	x	x	x	x



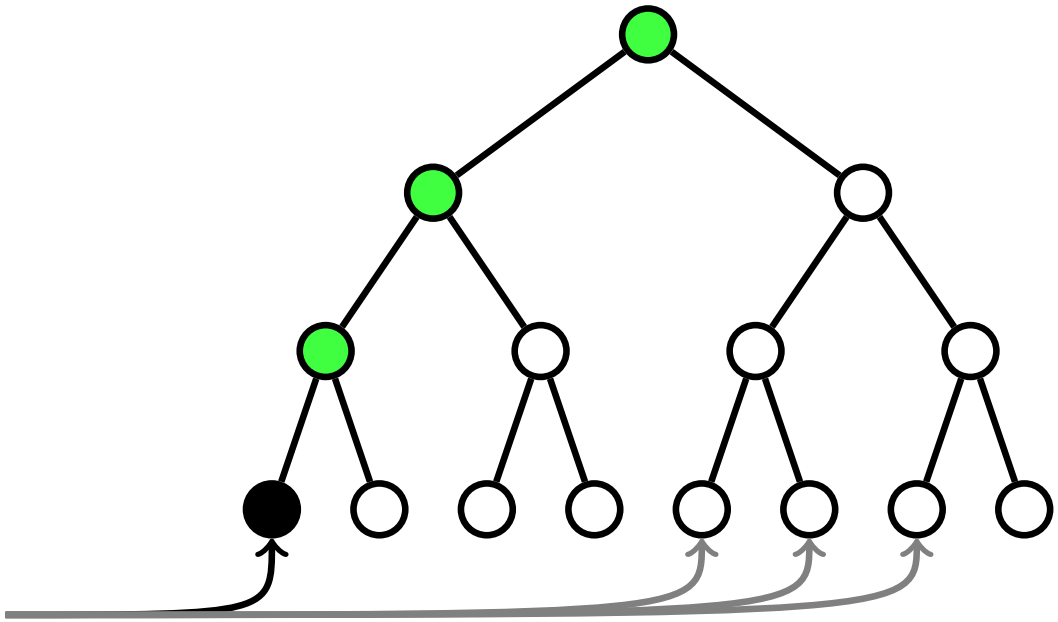
- Reads $R_1, \ldots, R_{10}$ match sequences $S_1, \ldots, S_4$

- Assume the reads in a metagenomic sample to be classified come from known sequences in a reference taxonomy
- The taxonomic annotation of the read at a certain node in the clade of the LCA in the reference taxonomy of the set of candidate sequences is correct if the candidate sequence that the read comes from lies in the clade of the node at which it is annotated

- What is the best indicator of classification error for the taxonomic annotation of metagenomic samples?

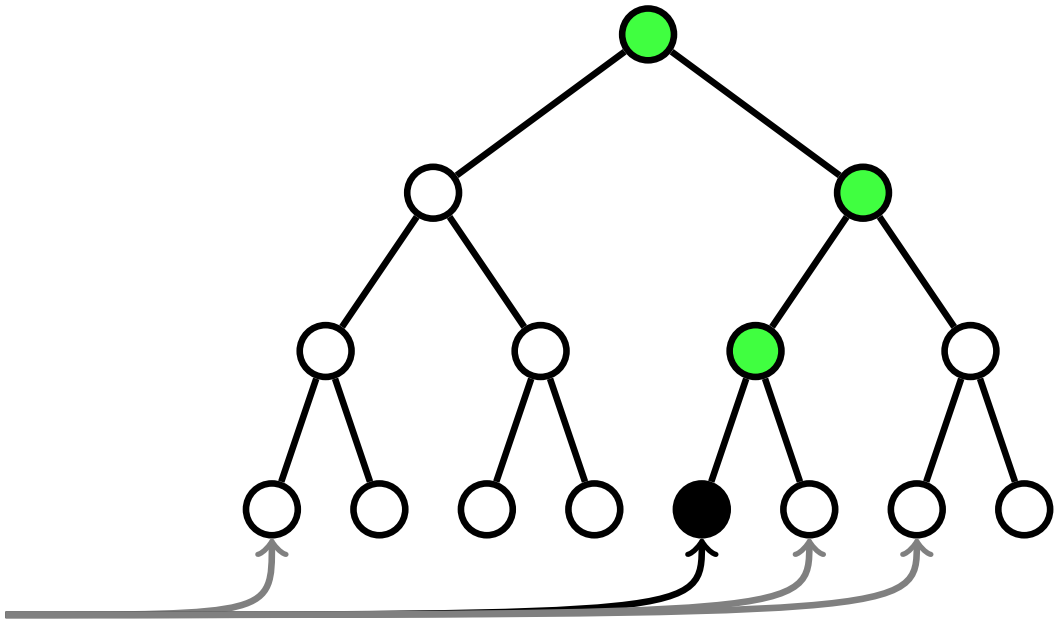
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT

ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT



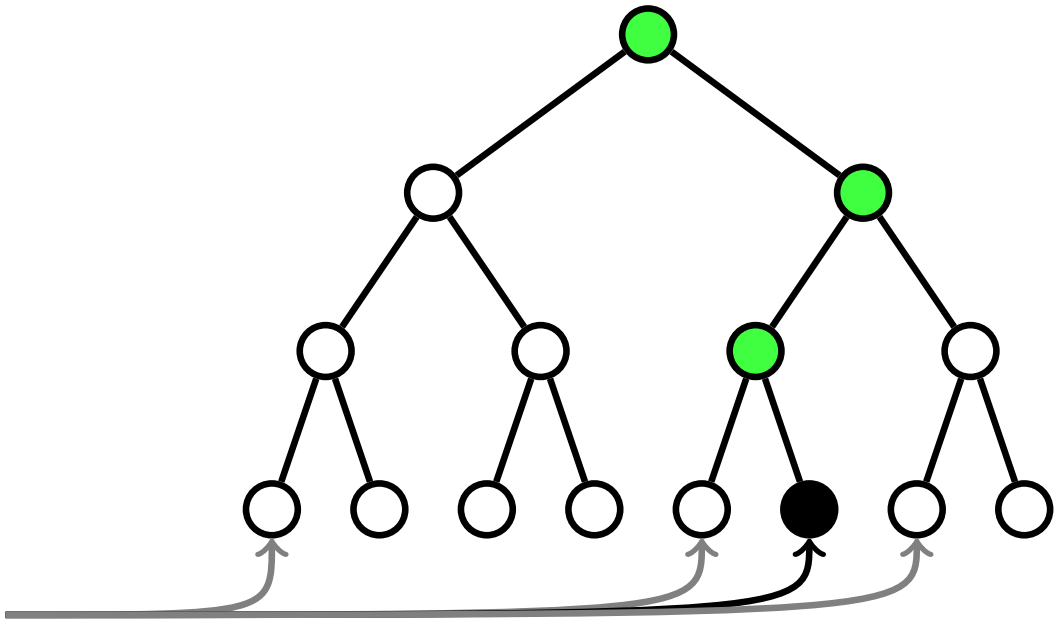
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT

ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT



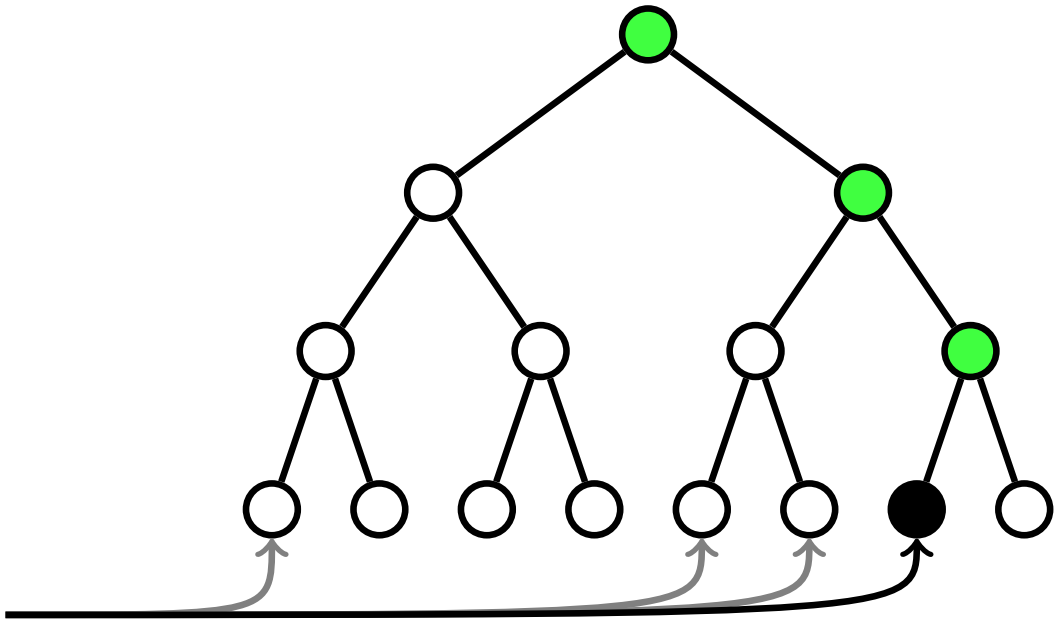
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT

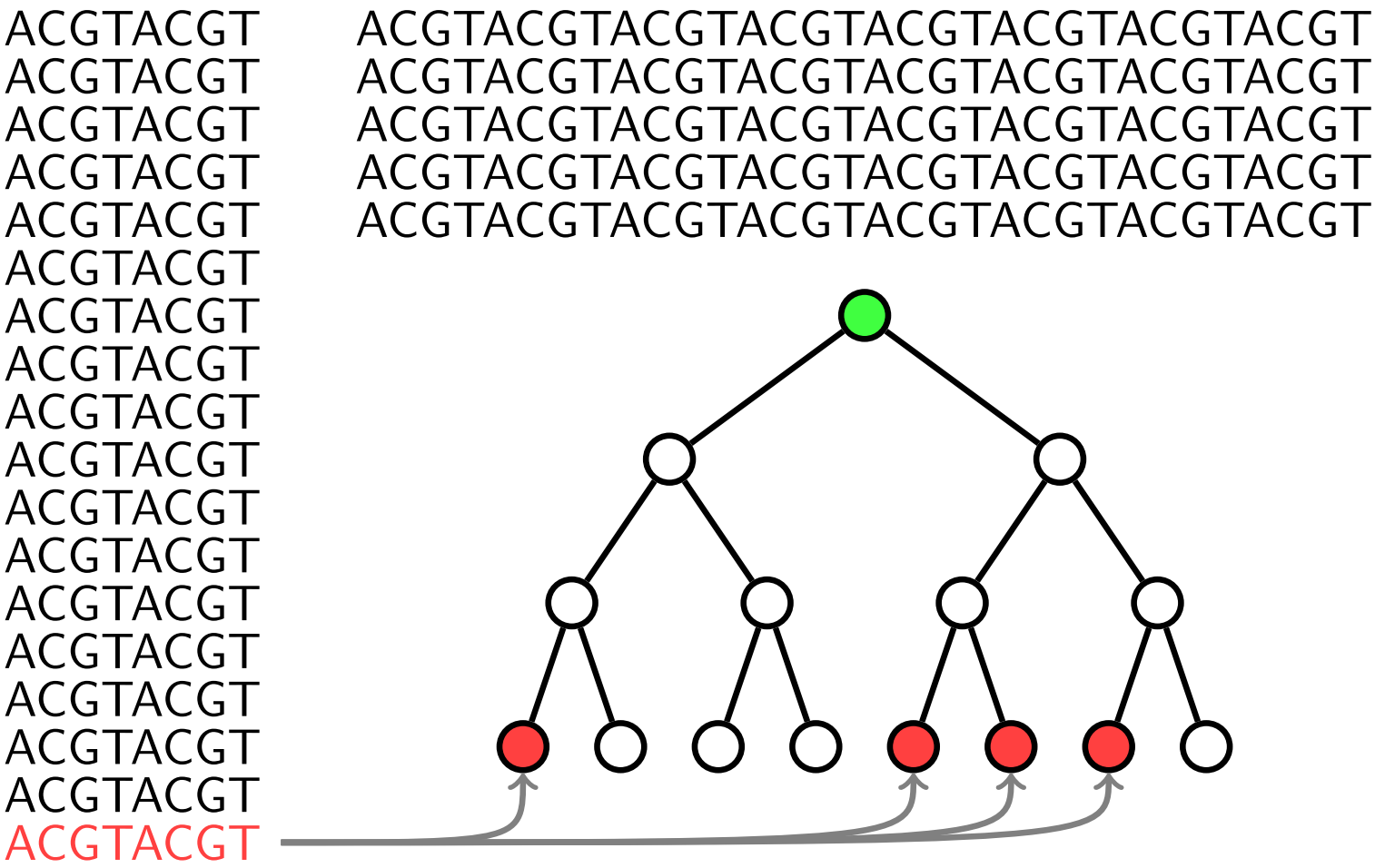
ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT



ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT

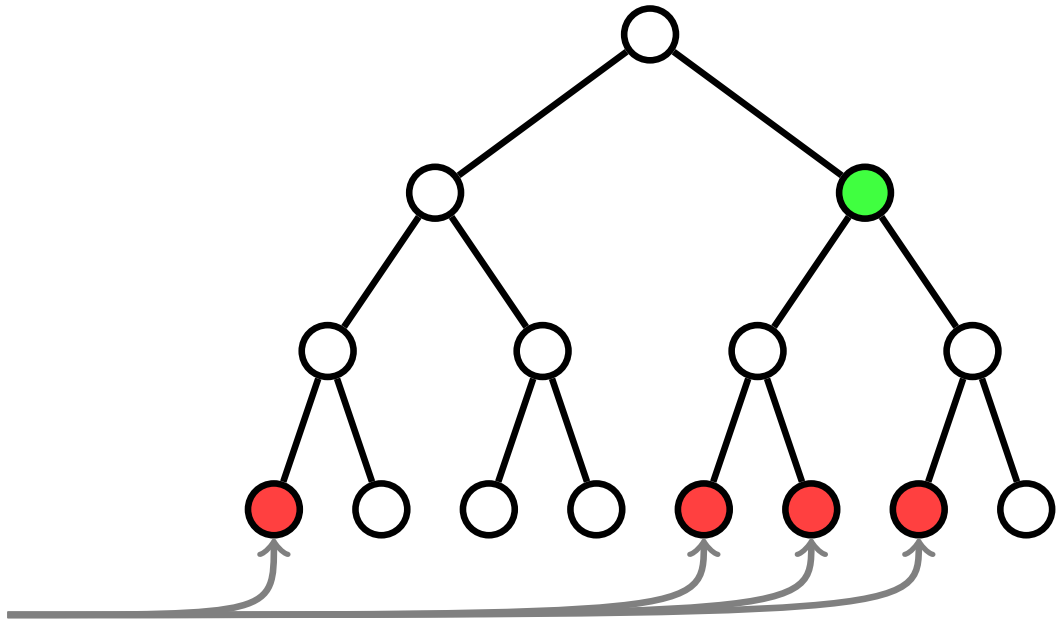
ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGT





- D. Huson and N. Weber. Microbial community analysis using MEGAN. In E. F. Delong, editor, *Methods in Enzymology*, volume 531, chapter 21, pages 465–485. Elsevier, 2013

ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT



- J. C. Clemente, J. Jansson, and G. Valiente. Flexible taxonomic assignment of ambiguous sequencing reads. *BMC Bioinformatics*, 12:8, 2011

Input

- A genomic reference S (set of sequences)
- A taxonomic reference T (tree) with a leaf set L, where each leaf in L has an associated known sequence of S
- A set R of sequence (short or long) reads
- A positive integer k

Output

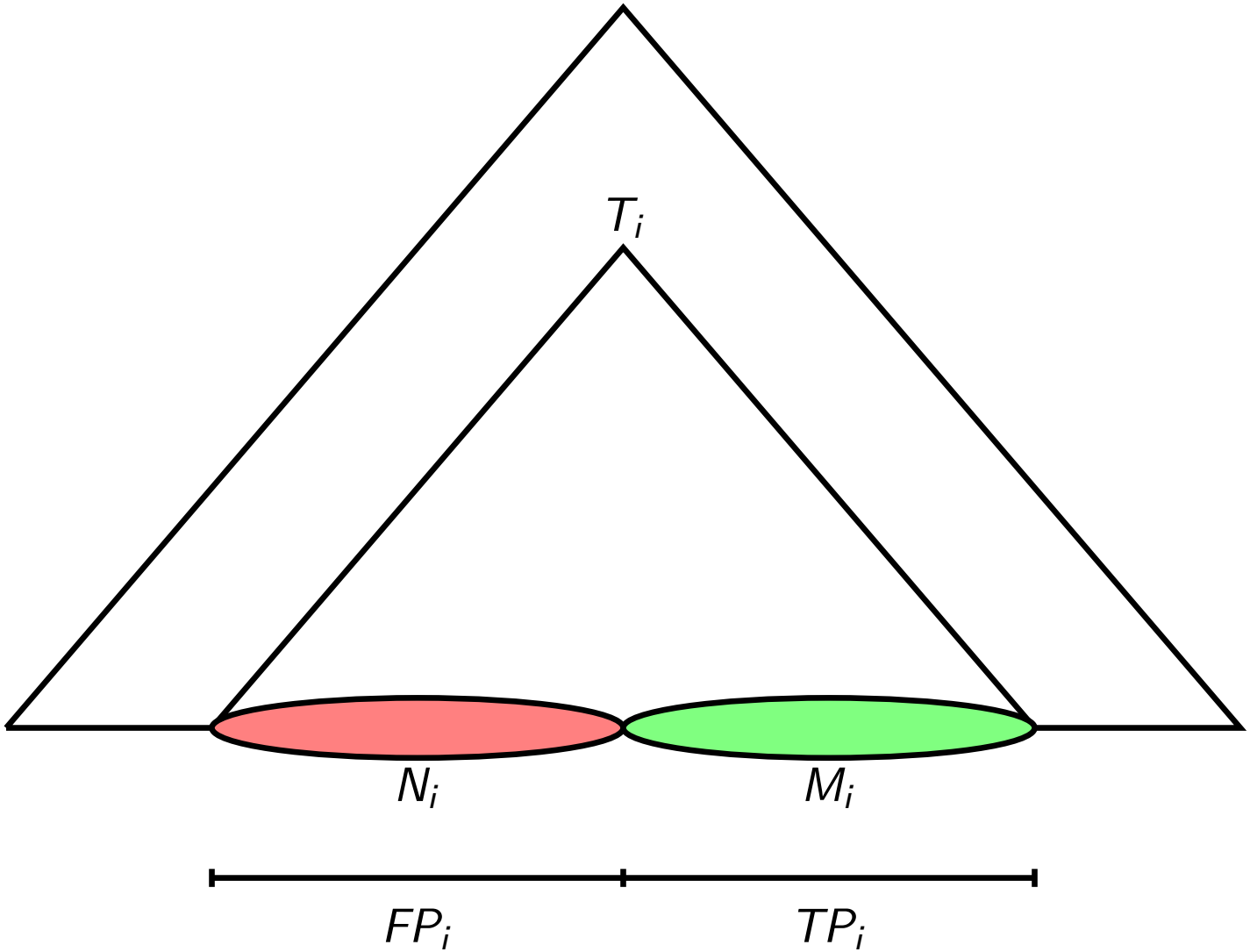
- For each read $R_i \in R$, a single node in T that represents in a "good" way the subset $M_i \subseteq L$ of hits or matches whose sequences contain a substring with at most k mismatches to R_i

Given a reference taxonomy T, a set R of sequence reads, and a threshold value k of sequence similarity,

- Let R_i be the ith read
- Let M_i be the leaves of T matching R_i with up to k mismatches
- Let T_i be the subtree of T rooted at the lowest common ancestor of M_i
- Let N_i be the leaves of T_i not matching R_i with up to k mismatches

For the ith read, the leaves of T_i can be partitioned in the following four subsets:

- $TP_i = M_i$ (true positives)
- $FP_i = N_i$ (false positives)
- $TN_i = \emptyset$ (true negatives)
- $FN_i = \emptyset$ (false negatives)


T_i
N_i
M_i
FP_i
TP_i

- Precision is the proportion of correctly labeled positive elements with respect to the total number of elements labeled positive (correctly or incorrectly labeled positive)

$$P = \frac{TP}{TP + FP}$$

- Recall is the proportion of correctly labeled positive elements with respect to the total number of positive elements (correctly labeled positive or incorrectly labeled negative)

$$R = \frac{TP}{TP + FN}$$

- The F-measure is the harmonic mean of precision and recall

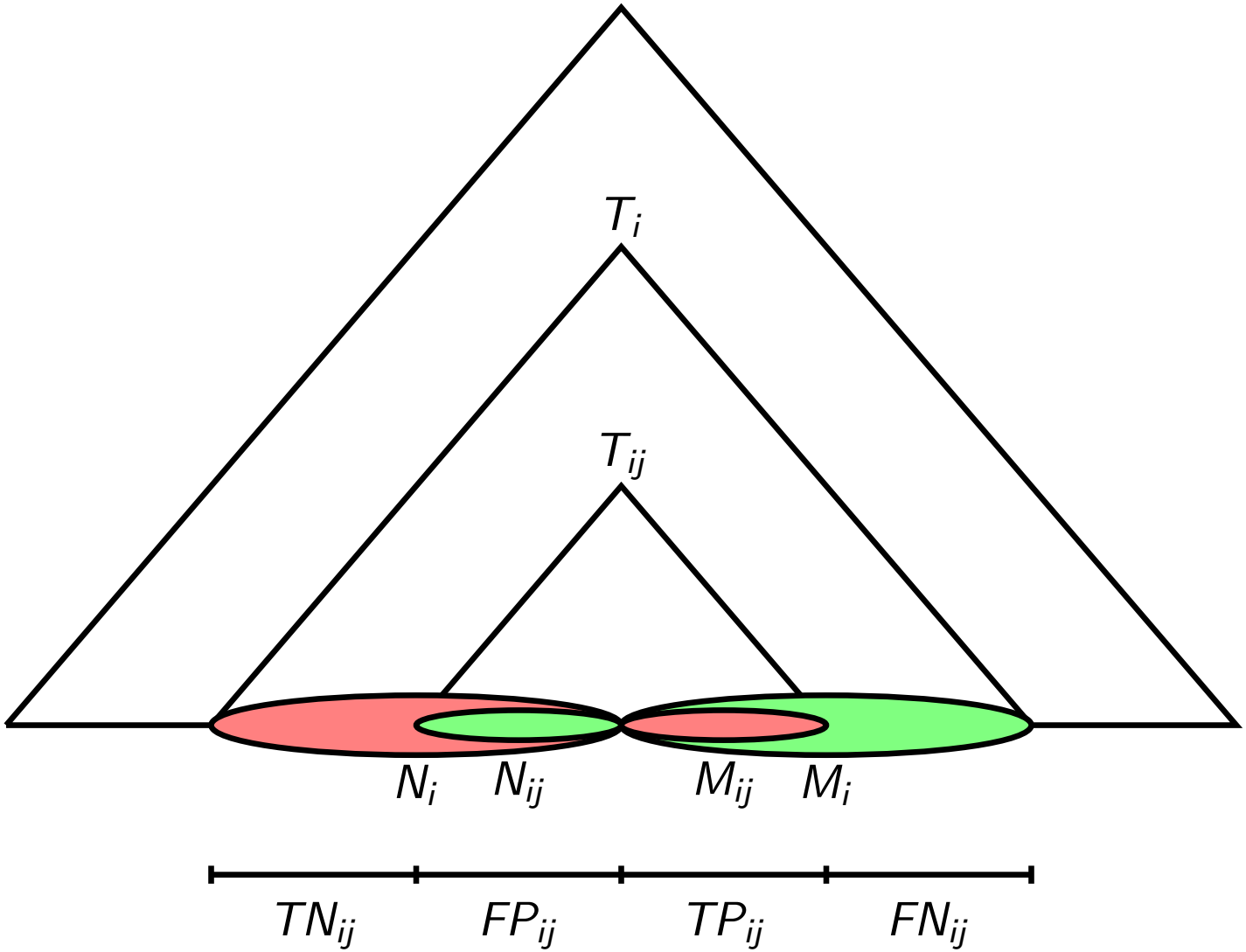
$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P + R}$$

Given a reference taxonomy T, a set R of sequence reads, and a threshold value k of sequence similarity,

- Let T_{ij} be the subtree of T rooted at the jth node of T_i
- Let M_{ij} be the leaves of T_{ij} matching R_i with up to k mismatches
- Let N_{ij} be the leaves of T_{ij} not matching R_i with up to k mismatches

For the ith read and the jth node of T_i, the leaves of T_i can be partitioned in the following four subsets:

- $TP_{ij} = M_{ij}$ (true positives)
- $FP_{ij} = N_{ij}$ (false positives)
- $TN_{ij} = N_i \setminus N_{ij}$ (true negatives)
- $FN_{ij} = M_i \setminus M_{ij}$ (false negatives)


T_i
T_{ij}
N_i
N_{ij}
M_{ij}
M_i
TN_{ij}
FP_{ij}
TP_{ij}
FN_{ij}

- The precision of classifying R_i as T_j is

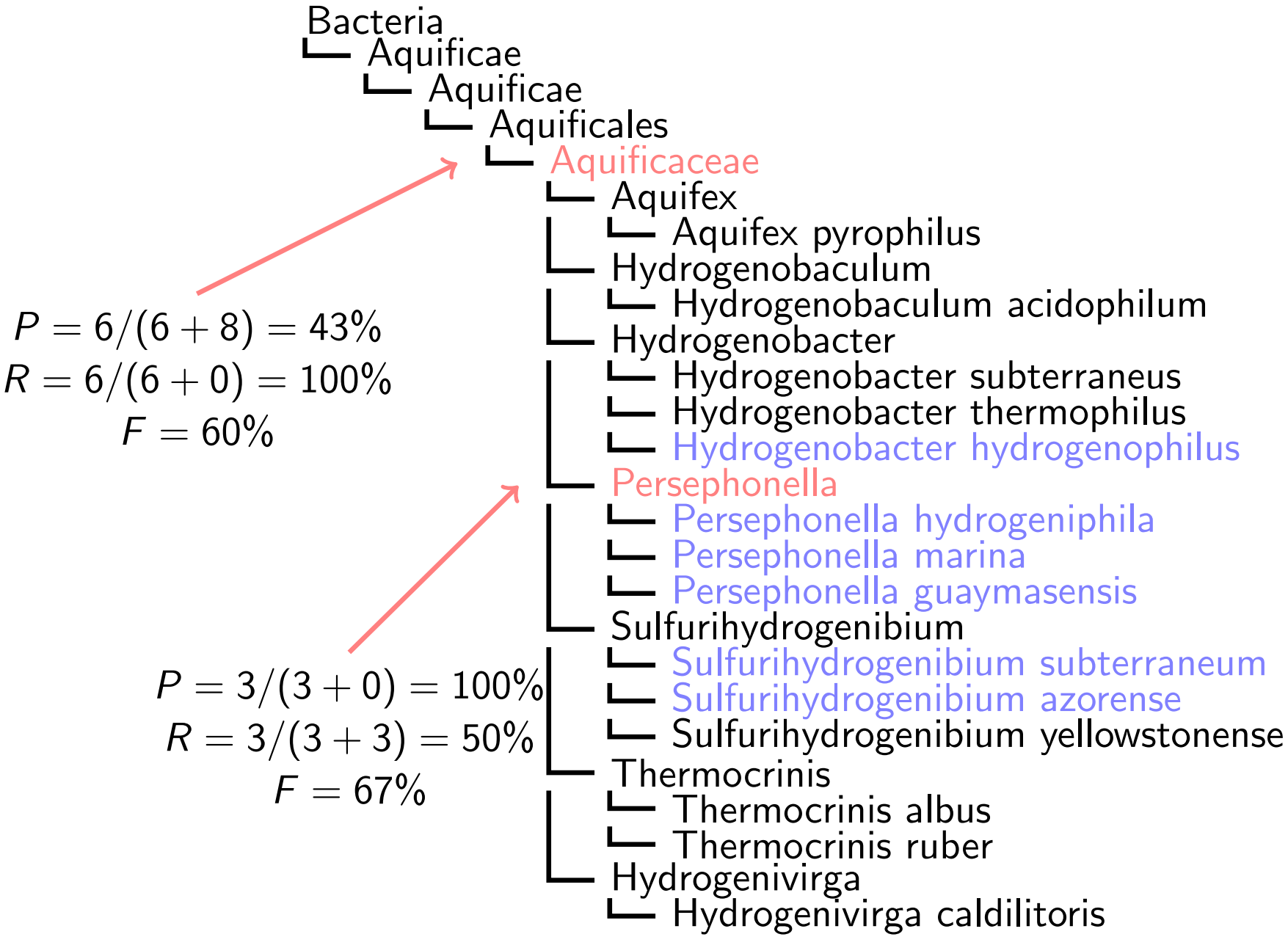
$$P_{ij} = \frac{|TP_{ij}|}{|TP_{ij}| + |FP_{ij}|}$$

- The recall of classifying R_i as T_j is

$$R_{ij} = \frac{|TP_{ij}|}{|TP_{ij}| + |FN_{ij}|}$$

- The combined F-measure of precision and recall is

$$F_{ij} = \frac{2}{\frac{1}{P_{ij}} + \frac{1}{R_{ij}}} = \frac{2P_{ij}R_{ij}}{P_{ij} + R_{ij}}$$


Bacteria
Aquificae
Aquificae
Aquificales
Aquificaceae
Aquifex
Aquifex pyrophilus
Hydrogenobaculum
Hydrogenobaculum acidophilum
Hydrogenobacter
Hydrogenobacter subterraneus
Hydrogenobacter thermophilus
Hydrogenobacter hydrogenophilus
Persephonella
Persephonella hydrogeniphila
Persephonella marina
Persephonella guaymasensis
Sulfurihydrogenibium
Sulfurihydrogenibium subterraneum
Sulfurihydrogenibium azorense
Sulfurihydrogenibium yellowstonense
Thermocrinis
Thermocrinis albus
Thermocrinis ruber
Hydrogenivirga
Hydrogenivirga caldilitoris
$P = 6/(6+8) = 43\%$
$R = 6/(6+0) = 100\%$
$F = 60\%$
$P = 3/(3+0) = 100\%$
$R = 3/(3+3) = 50\%$
$F = 67\%$

- The combined *F*-measure of precision and recall is

$$F_{ij} = \frac{2P_{ij}R_{ij}}{P_{ij} + R_{ij}} = \frac{2|TP_{ij}|}{|FN_{ij}| + |FP_{ij}| + 2|TP_{ij}|}$$

- The penalty score of assigning R_i to T_j is

$$PS_{ij} = q\frac{|FN_{ij}|}{|TP_{ij}|} + (1-q)\frac{|FP_{ij}|}{|TP_{ij}|}$$

The parameter q takes values in the range $[0, 1]$ and determines how close to the LCA or to the leaves the assignment shall be

- When $q = 0$, each read R_i is assigned to a matching leaf
- When $q = 0.5$, each read R_i is assigned to the node that maximizes the *F*-measure
- When $q = 1$, each read R_i is assigned to the LCA of the matching leaves M_i

The penalty score is a generalization of the F-measure

- For $q = 0.5$, the node m that minimizes the penalty score is

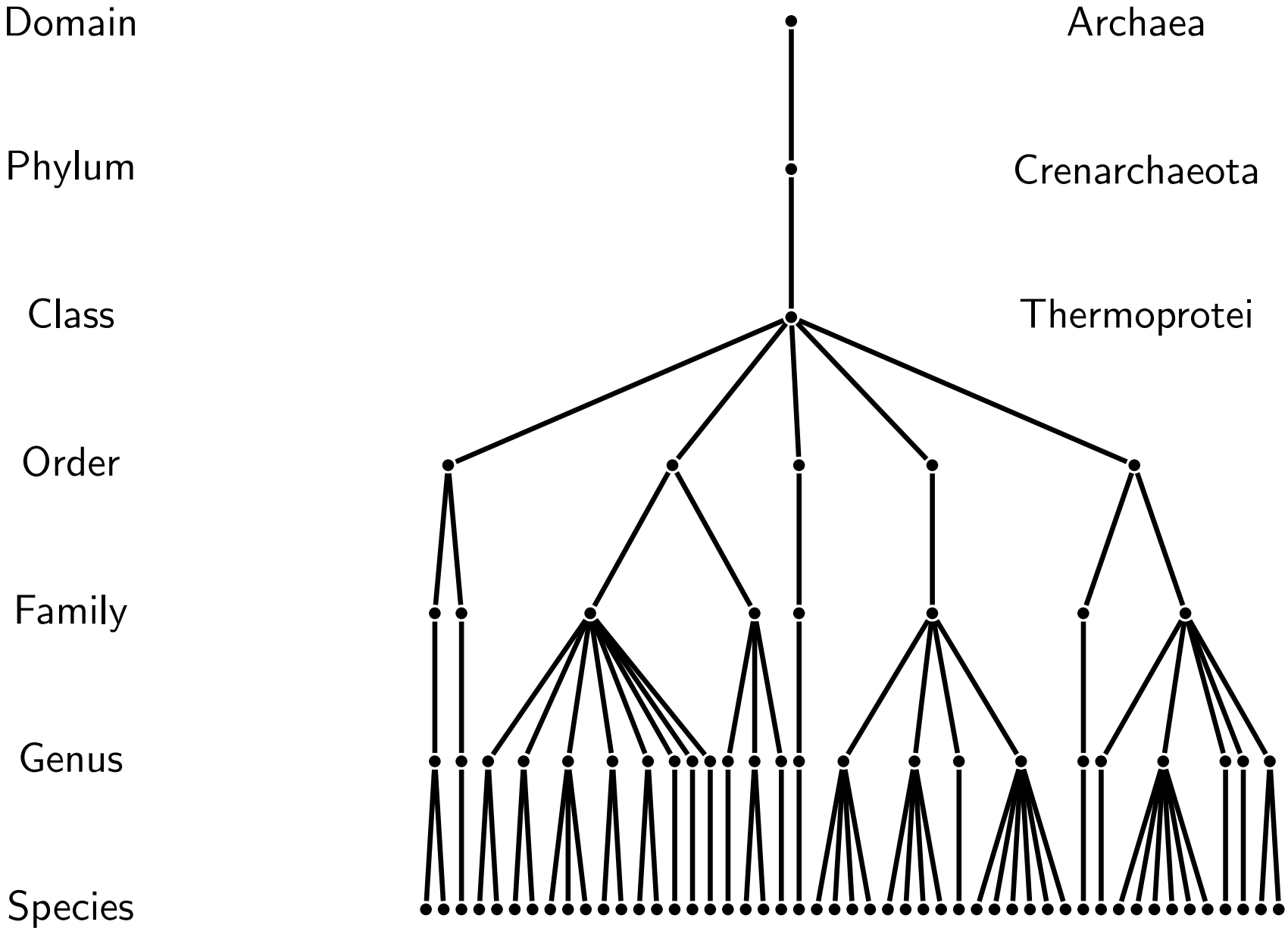
$$\begin{aligned}
&\arg\min_m(|FN_{im}|/|TP_{im}| + |FP_{im}|/|TP_{im}|) \\
&= \arg\min_m((|FN_{im}| + |FP_{im}|)/|TP_{im}|) \\
&= \arg\min_m((|FN_{im}| + |FP_{im}|)/2|TP_{im}|) \\
&= \arg\min_m(((|FN_{im}| + |FP_{im}|)/2|TP_{im}|) + 1) \\
&= \arg\min_m((|FN_{im}| + |FP_{im}| + 2|TP_{im}|)/2|TP_{im}|) \\
&= \arg\max_m(2|TP_{im}|/(|FN_{im}| + |FP_{im}| + 2|TP_{im}|))
\end{aligned}$$

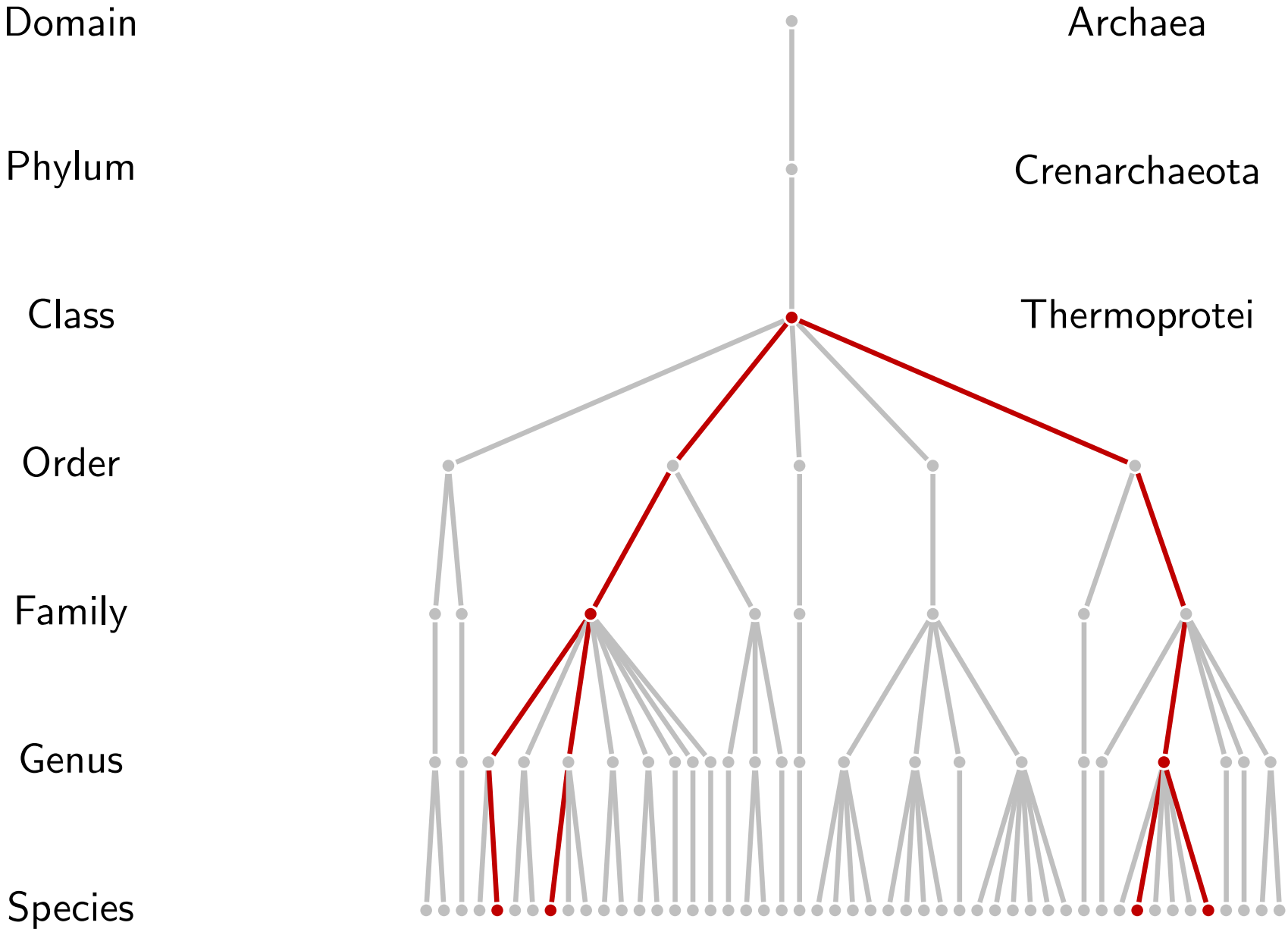
The node that minimizes the penalty score is the same node that would maximize the F-measure

- Given a set $M_i \subseteq L$ of hits and the subtree T_i of T rooted at the LCA of M_i, the penalty scores $PS_{i,j}$ for every node j in T_i can be obtained in $O(|T_i|)$ total time
- Given a set $M_i \subseteq L$ of hits and the subtree T_i of T rooted at the LCA of M_i, the penalty scores $PS_{i,j}$ for every node j in T_i can be obtained in $O(|M_i|)$ total time after $O(|T|)$ time preprocessing
- Any node j in T_i is called relevant if it is a leaf in M_i or the LCA of two or more leaves in M_i
- For each node j in T_i there exists a relevant node j' such that $PS_{i,j'} \leqslant PS_{i,j}$

- J. Fischer and D. H. Huson. New common ancestor problems in trees and directed acyclic graphs. *Information Processing Letters*, 110(8–9):331–335, 2010
- J. C. Clemente, J. Jansson, and G. Valiente. Flexible taxonomic assignment of ambiguous sequencing reads. *BMC Bioinformatics*, 12:8, 2011

```
domain Archaea
  phylum Crenarchaeota
    class Thermoprotei
      order Desulfurococcales
        family Desulfurococcaceae
          genus Aeropyrum
***         species Aeropyrum pernix
          genus Ignicoccus
***         species Ignicoccus islandicus
      order Thermoproteales
        family Thermoproteaceae
          genus Pyrobaculum
***         species Pyrobaculum oguniense
***         species Pyrobaculum organotrophum
      order Acidilobales
        family Acidilobaceae
          genus Acidilobus
            species Acidilobus aceticus
            species Acidilobus saccharovorans
        family Caldisphaeraceae
          genus Caldisphaera
            species Caldisphaera lagunensis
      order Desulfurococcales
        family Desulfurococcaceae
          genus Aeropyrum
            species Aeropyrum camini
***         species Aeropyrum pernix
          genus Desulfurococcus
            species Desulfurococcus amylolyticus
            species Desulfurococcus fermentans
          genus Ignicoccus
***         species Ignicoccus islandicus
            species Ignicoccus pacificus
            species Ignicoccus hospitalis
          genus Ignisphaera
            species Ignisphaera aggregans
            species Ignisphaera aggregans
          genus Staphylothermus
            species Staphylothermus marinus
            species Staphylothermus hellenicus
          genus Stetteria
            species Stetteria hydrogenophila
          genus Thermodiscus
            species Thermodiscus maritimus
          genus Thermosphaera
            species Thermosphaera aggregans
        family Pyrodictiaceae
          genus Hyperthermus
            species Hyperthermus butylicus
          genus Pyrodictium
            species Pyrodictium abyssi
            species Pyrodictium occultum
          genus Pyrolobus
            species Pyrolobus fumarii
      order Fervidicoccales
        family Fervidicoccaceae
          genus Fervidicoccus
            species Fervidicoccus fontis
      order Sulfolobales
        family Sulfolobaceae
          genus Acidianus
            species Acidianus brierleyi
            species Acidianus infernus
            species Acidianus ambivalens
            species Acidianus sulfidivorans
          genus Metallosphaera
            species Metallosphaera sedula
            species Metallosphaera hakonensis
            species Metallosphaera prunae
            species Metallosphaera cuprina
          genus Stygiolobus
            species Stygiolobus azoricus
          genus Sulfolobus
            species Sulfolobus yangmingensis
            species Sulfolobus tokodaii
            species Sulfolobus acidocaldarius
            species Sulfolobus solfataricus
            species Sulfolobus metallicus
            species Sulfolobus shibatae
      order Thermoproteales
        family Thermofilaceae
          genus Thermofilum
            species Thermofilum pendens
        family Thermoproteaceae
          genus Caldivirga
            species Caldivirga maquilingensis
          genus Pyrobaculum
            species Pyrobaculum calidifontis
***         species Pyrobaculum oguniense
            species Pyrobaculum aerophilum
            species Pyrobaculum islandicum
            species Pyrobaculum arsenaticum
***         species Pyrobaculum organotrophum
          genus Thermocladium
            species Thermocladium modestius
          genus Thermoproteus
            species Thermoproteus tenax
          genus Vulcanisaeta
            species Vulcanisaeta distributa
            species Vulcanisaeta souniana
```


Domain
Archaea
Phylum
Crenarchaeota
Class
Thermoprotei
Order
Family
Genus
Species


Domain
Archaea
Phylum
Crenarchaeota
Class
Thermoprotei
Order
Family
Genus
Species