Cold Spring Harbor Laboratory
DNA Learning Center

Barcoding Bioinformatics Part III

Jason Williams
Cold Spring Harbor Laboratory, DNA Learning Center
williams@cshl.edu
@JasonWilliamsNY

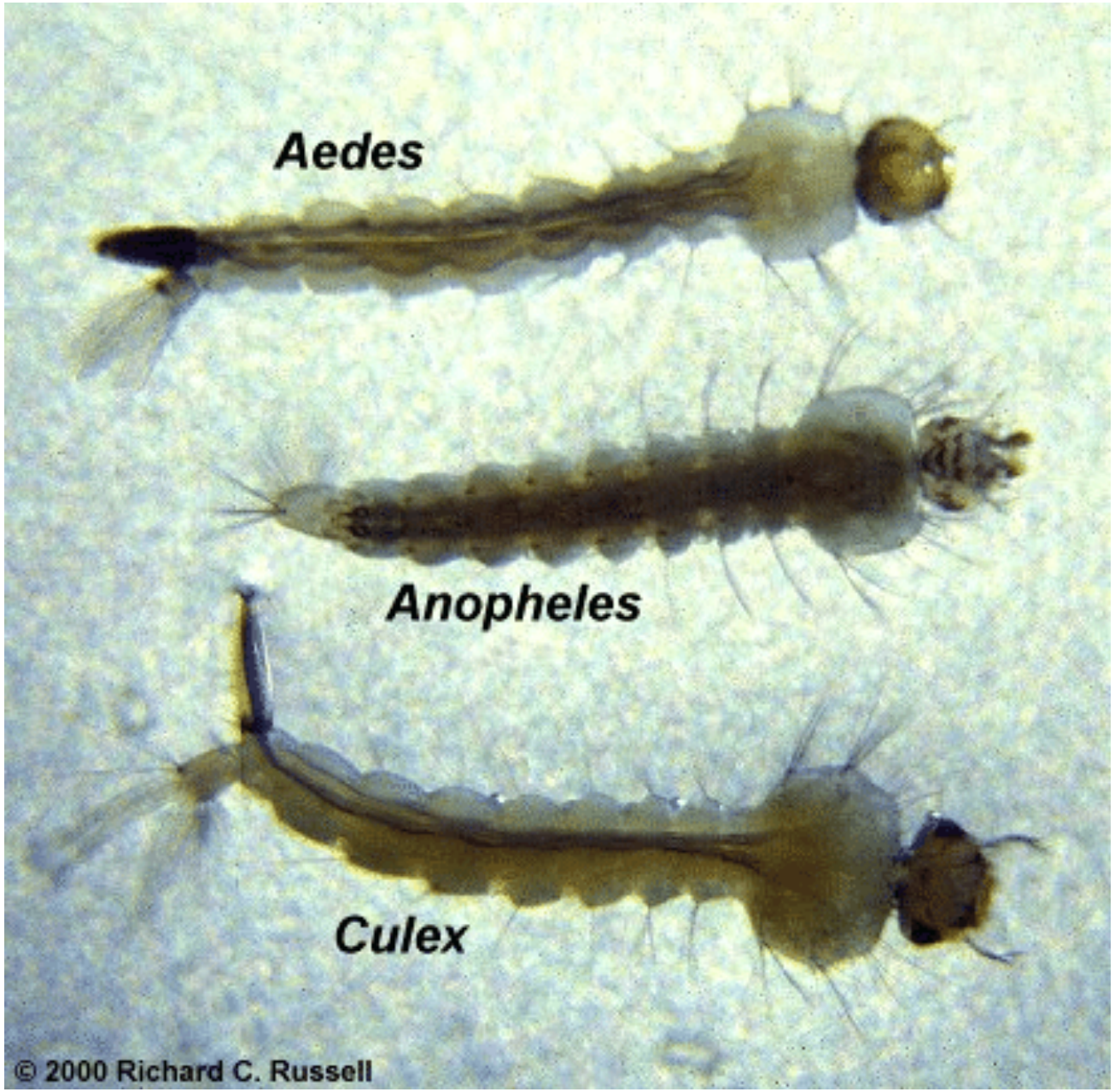
Barcoding Bioinformatics Part III

(Sequence alignment and phylogenetics)

Steps for today's session

- Recap on our experimental dataset
- Review of BLAST
- Introduction to multiple alignment
- Introduction to phylogenetic trees

Recap of the dataset


Aedes
Anopheles
Culex
© 2000 Richard C. Russell

Aedes adult

By Muhammad Mahdi Karim - Own work, GFDL 1.2, https://commons.wikimedia.org/w/index.php?curid=11185617

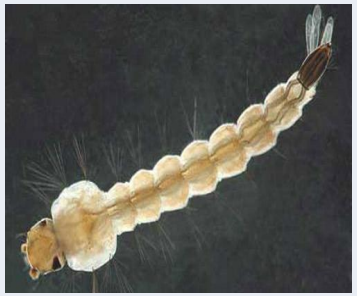
Anopheles adult

By Jim Gathany - (PHIL), ID #5814. https://commons.wikimedia.org/w/index.php?curid=799284

Culex adult

By Muhammad Mahdi Karim - Own work, GFDL 1.2, https://commons.wikimedia.org/w/index.php?curid=7673048

Aedes larva



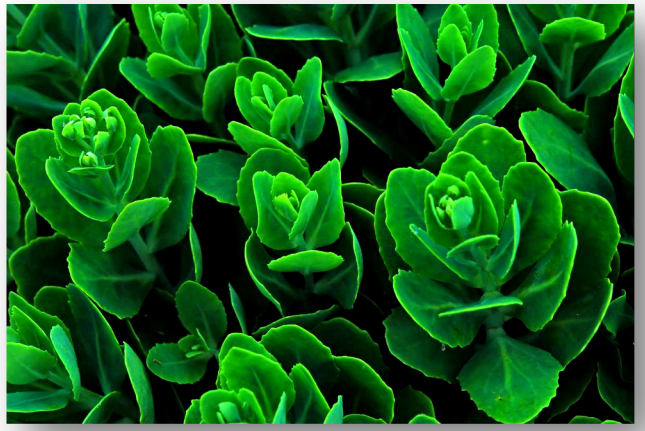
Photograph by Michele M. Cutwa, University of Florida.

Anopheles larva

Culex larva

Photograph by Michelle Cutwa-Francis, University of Florida.

Steps to DNA Barcoding



Organism is sampled

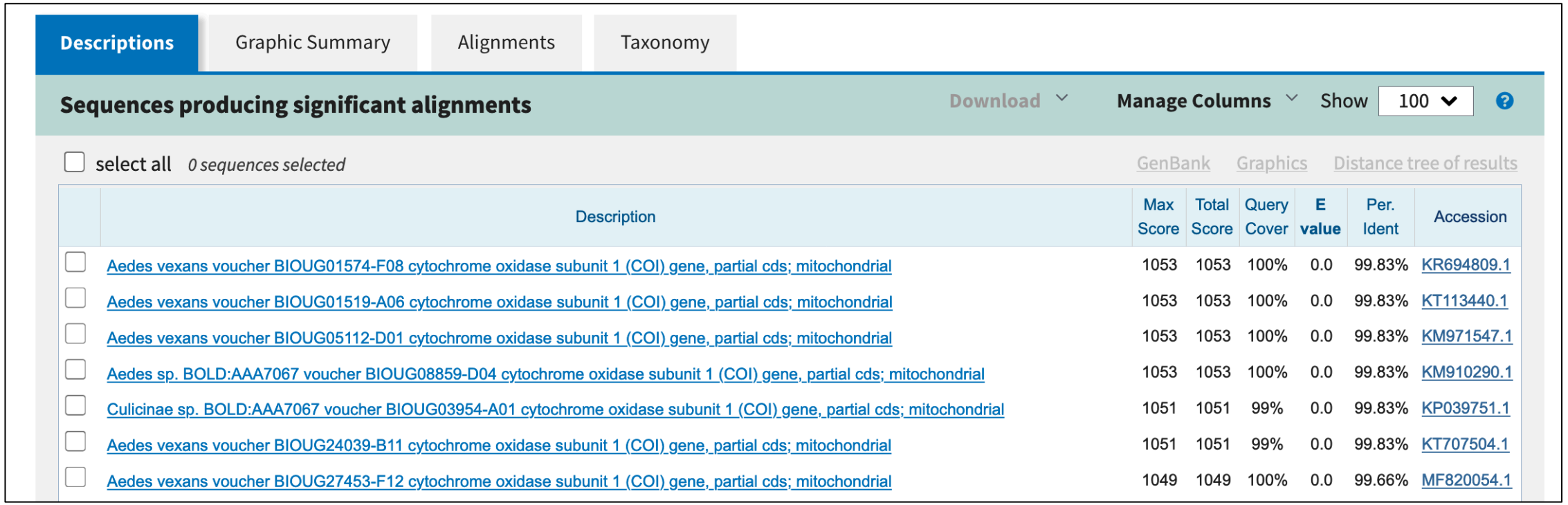
DNA is extracted

“Barcode” amplified

ACGAGTCGGTAGCTGCCCTCTGACTGCATCGAA
TTGCTCCCCTACTACGTGCTATATGCGCTTACGAT
CGTACGAAGATTTATAGAATGCTGCTACTGCTCC
CTTATTCGATAACTAGCTCGATTATAGCTACGATG

Sequenced DNA is compared with DNA in a barcode database

Let’s do a BLAST



Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Manage Columns Show 100

select all *0 sequences selected* GenBank Graphics Distance tree of results

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Aedes vexans voucher BIOUG01574-F08 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	1053	1053	100%	0.0	99.83%	KR694809.1
Aedes vexans voucher BIOUG01519-A06 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	1053	1053	100%	0.0	99.83%	KT113440.1
Aedes vexans voucher BIOUG05112-D01 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	1053	1053	100%	0.0	99.83%	KM971547.1
Aedes sp. BOLD:AAA7067 voucher BIOUG08859-D04 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	1053	1053	100%	0.0	99.83%	KM910290.1
Culicinae sp. BOLD:AAA7067 voucher BIOUG03954-A01 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	1051	1051	99%	0.0	99.83%	KP039751.1
Aedes vexans voucher BIOUG24039-B11 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	1051	1051	99%	0.0	99.83%	KT707504.1
Aedes vexans voucher BIOUG27453-F12 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	1049	1049	100%	0.0	99.66%	MF820054.1

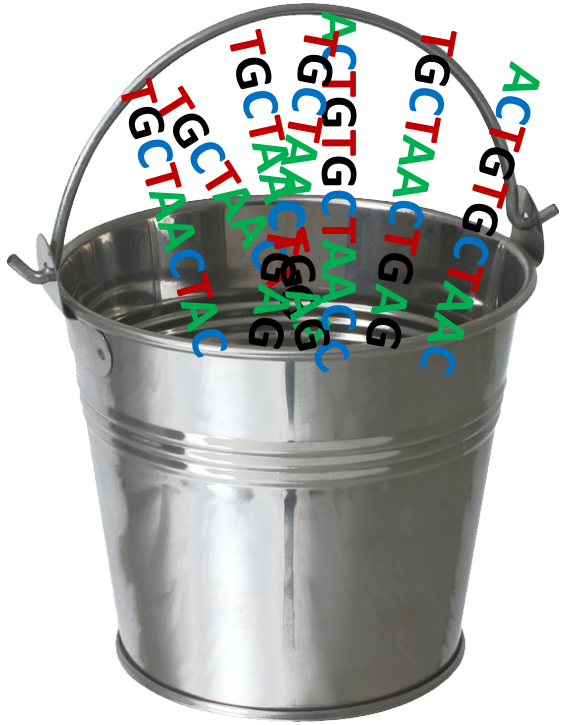
Basic Local Alignment Search Tool

- An algorithm for searching a database of sequences
- “Google for DNA” (although works with any biological sequence, and started before Google ~1990 vs 1998)
- NCBI is the most popular interface, but this is software that can be run anywhere (including Subway)

BLAST algorithm analogy

Query sequence

ACTGACATCGGGGTGCTACG



Database

Where in the DNA should we look for the "Barcode"?

Where in the DNA should we look?

Humans share greater than 99% of their DNA with other humans

Photo Credit:
https://www.broadinstitute.org/files/styles/landing_page/public/generic-pages/images/circle/MPG-mosaic_385x300.png?itok=B5zrVgYz

Where in the DNA should we look?

Photo credit:
https://www.cnbc.com/2015/09/02/can-you-find-the-forgery.html

Where in the DNA should we look?

Photo credit:
https://www.cnbc.com/2015/09/02/can-you-find-the-forgery.html

Where in the DNA should we look?

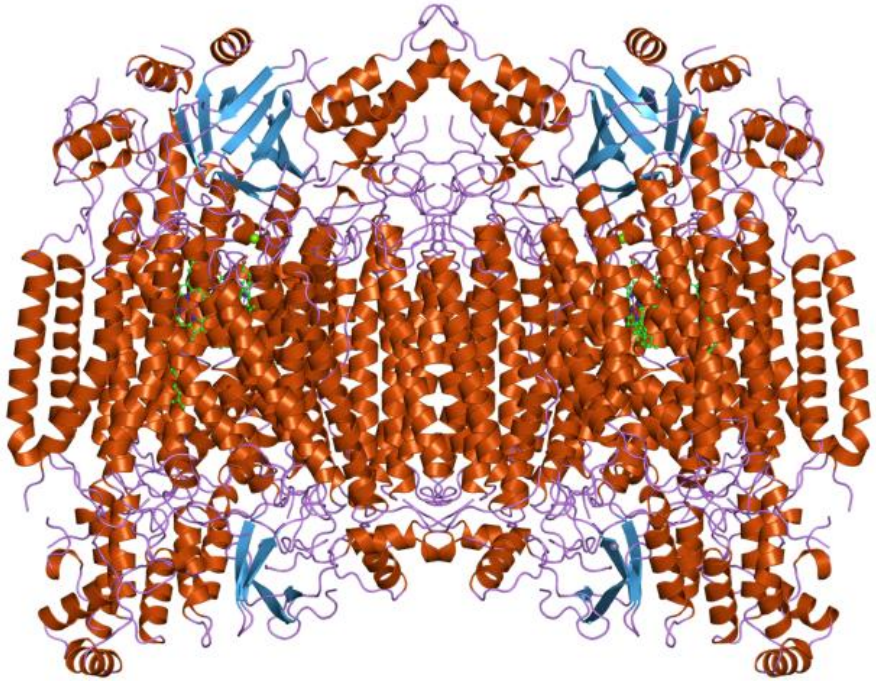
>Human Tubulin –Black?
GCAGGTTCTCTTACATCGACCGCCTAAGAGTCGCGCTGTAAGAAGCAACAACCTCTCCTCTTCGTCTCCG
CCATCAGCTCGGCAGTCGCGAAGCAGCAACCATGCGTGAGTGCATCTCCATCCACGTTGGCCAGGCTGGT
GTCCAGATTGGCAATGCCTGCTGGGAGCTCTACTGCCTGGAACACGGCATCCAGCCCGATGGCCAGATGC
CAAGTGACAAGACCATTGGGGGAGGAGATGATTCCTTCAACACCTTCTTCAGTGAGACGGGGGCTGGCAA
GCATGTGCCCCGGGCAGTGTTTGTAGACTTGGAACCCACAGTCATTGATGAAGTTCGCACTGGCACCTAC
CGCCAGCTCTTCCACCCTGAGCAACTTATCACAGGCAAAGAAGATGCTGCCAATAACTATGCCCGAGGGC
ACTACACCATTGGCAAGGAGATCATTGACCTCGTGTTGGACCGAATTCGCAAGCTGGCCGACCAGTGCAC

>Human Tubulin –White?
GCAGGTTCTCTTACATCGACCGCCTAAGAGTCGCGCTGTAAGAAGCAACAACCTCTCCTCTTCGTCTCCG
CCATCAGCTCGGCAGTCGCGAAGCAGCAACCATGCGTGAGTGCATCTCCATCCACGTTGGCCAGGCTGGT
GTCCAGATTGGCAATGCCTGCTGGGAGCTCTACTGCCTGGAACACGGCATCCAGCCCGATGGCCAGATGC
CAAGTGACAAGACCATTGGGGGAGGAGATGATTCCTTCAACACCTTCTTCAGTGAGACGGGGGCTGGCAA
GCATGTGCCCCGGGCAGTGTTTGTAGACTTGGAACCCACAGTCATTGATGAAGTTCGCACTGGCACCTAC
CGCCAGCTCTTCCACCCTGAGCAACTTATCACAGGCAAAGAAGATGCTGCCAATAACTATGCCCGAGGGC
ACTACACCATTGGCAAGGAGATCATTGACCTCGTGTTGGACCGAATTCGCAAGCTGGCCGACCAGTGCAC

>Human Tubulin –Asian?
GCAGGTTCTCTTACATCGACCGCCTAAGAGTCGCGCTGTAAGAAGCAACAACCTCTCCTCTTCGTCTCCG
CCATCAGCTCGGCAGTCGCGAAGCAGCAACCATGCGTGAGTGCATCTCCATCCACGTTGGCCAGGCTGGT
GTCCAGATTGGCAATGCCTGCTGGGAGCTCTACTGCCTGGAACACGGCATCCAGCCCGATGGCCAGATGC
CAAGTGACAAGACCATTGGGGGAGGAGATGATTCCTTCAACACCTTCTTCAGTGAGACGGGGGCTGGCAA
GCATGTGCCCCGGGCAGTGTTTGTAGACTTGGAACCCACAGTCATTGATGAAGTTCGCACTGGCACCTAC
CGCCAGCTCTTCCACCCTGAGCAACTTATCACAGGCAAAGAAGATGCTGCCAATAACTATGCCCGAGGGC
ACTACACCATTGGCAAGGAGATCATTGACCTCGTGTTGGACCGAATTCGCAAGCTGGCCGACCAGTGCAC

Almost everywhere we look in our (human) genome, we all look the same

Cytochrome c oxidase I (COI)



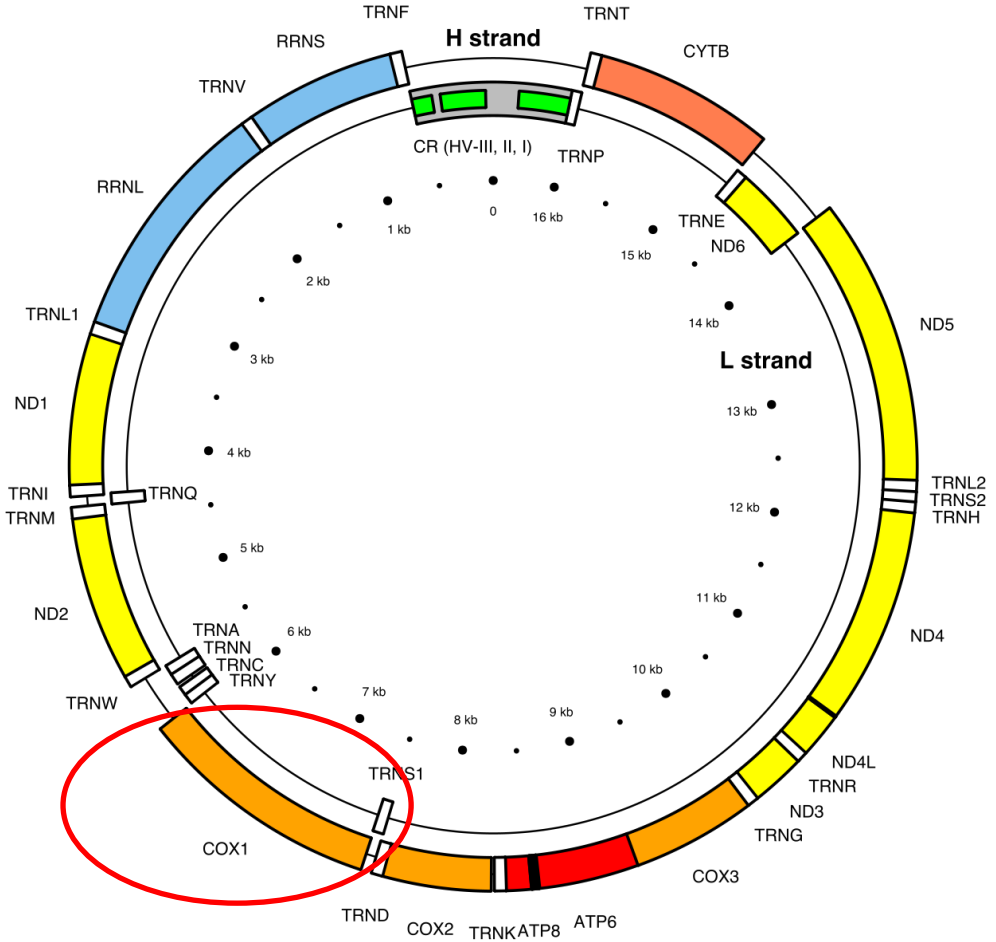


Photo credit:
Structure:
https://en.wikipedia.org/wiki/Cytochrome_c_oxidase_subunit_I#/media/File:PDB_1occ_EBI.jpg
Genome map:
Emmanuel Douzery;
https://en.wikipedia.org/wiki/Cytochrome_c_oxidase_subunit_I#/media/File:Map_of_the_human_mitochondrial_genome.svg

Choosing a barcoding locus

There are many criteria that go in to selecting an appropriate locus (location in the genome) that can serve as a barcode.

Three of them include:

- Universality
- Discrimination
- Robustness

Universality

Since barcoding protocols (typically) amplify a region of DNA by PCR, you need to choose a DNA sequence that every species has

Universality

Since barcoding protocols (typically) amplify a region of DNA by PCR, you need to choose a DNA sequence that every species has

Universality

Since barcoding protocols (typically) amplify a region of DNA by PCR, you need to choose a DNA sequence that every species has

Carnivores
- Cat
  3.1.3.1
  3.1.2.1
- Dog
  3.1.4.2
  3.1.4.3

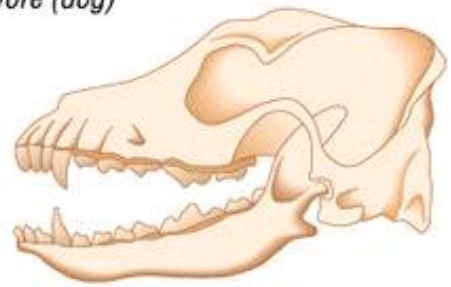
Omnivores
- Pig
  3.1.4.3
  3.1.4.3
- Human
  2.1.2.3
  2.1.2.3

Herbivores
- Cow
  0.0.3.3
  4.0.3.3
- Horse
  3.1.3/4.3
  3.1.3.3
- Rabbit
  2.0.3.3
  1.0.2.3
- Sheep
  0.0.3.3
  3.0.3.3

herbivore (sheep)

dental formula $i\frac{0}{3}\ c\frac{0}{0}\ pm\frac{3}{3}\ m\frac{3}{3}$

carnivore (dog)



dental formula $i\frac{3}{3}\ c\frac{1}{1}\ pm\frac{4}{4}\ m\frac{2}{3}$

Photo credit:
Dental formula
https://slideplayer.com/slide/10972461/
Animal diagrams
https://vet-science.blogspot.com/2012/01/dentition-in-sheep-and-goat.html

Discrimination

Barcoding regions must be different for each species. Ideally you are looking for a single DNA locus which differs in each species

Discrimination

Barcoding regions must be different for each species. Ideally you are looking for a single DNA locus which differs in each species

Photo credit
Human hair
https://lewigs.com/human-hair-color-101/

Discrimination

Barcoding regions must be different for each species. Ideally you are looking for a single DNA locus which differs in each species

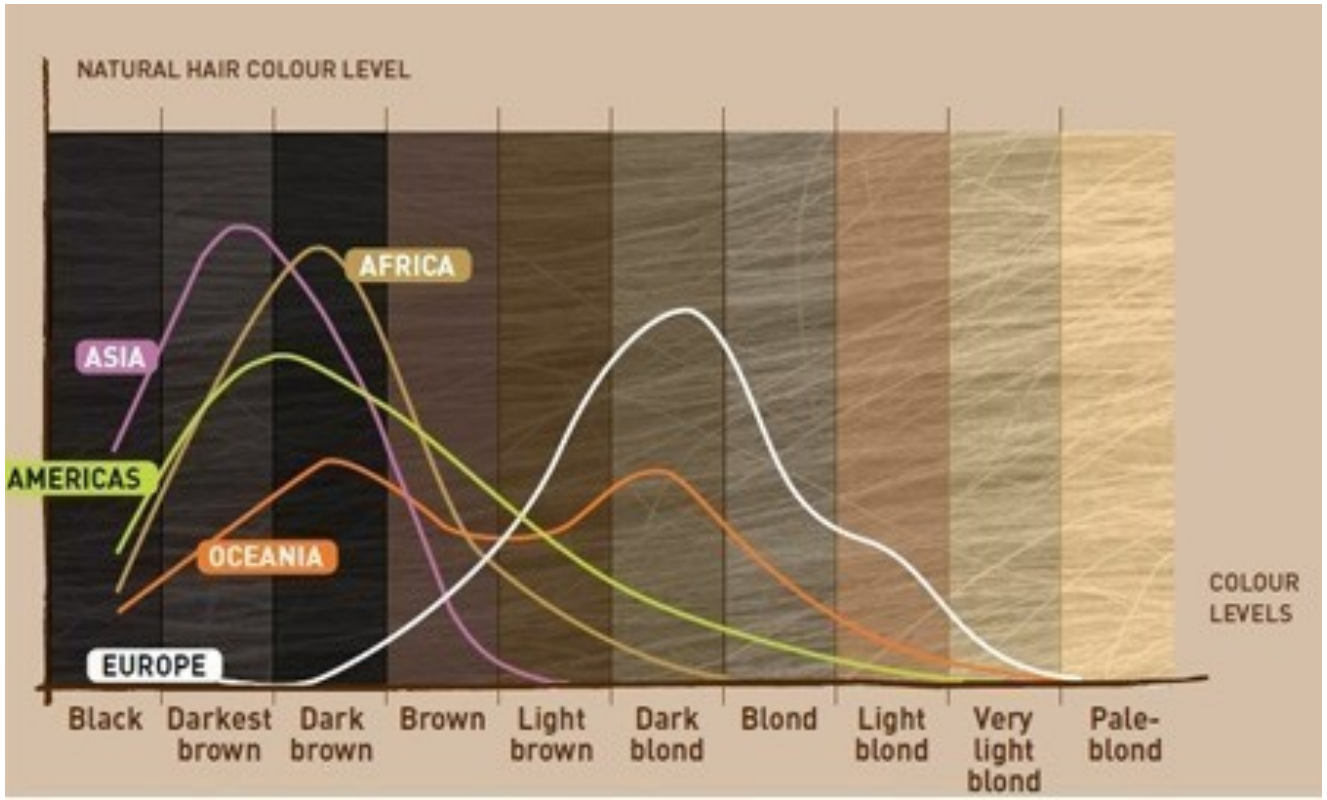
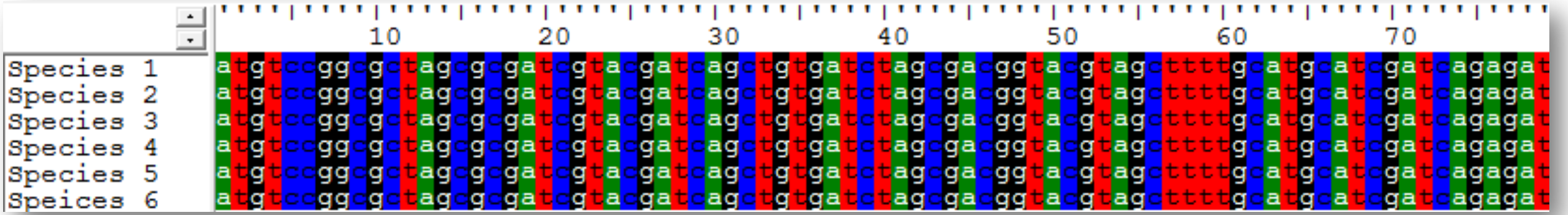


Photo credit
http://loreal-dam-videos-corp-en-cdn.brainsonic.com/corpen/20160330pm/20160330-170533-c4bb4cb7/picture_photo_3eadc4.jpg

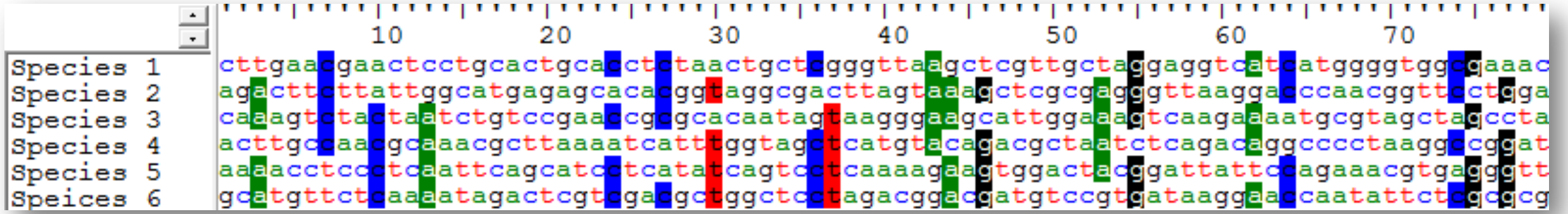
Discrimination (but not too much)

Fail: Sequence is completely conserved, good for PCR, but uninformative as barcode



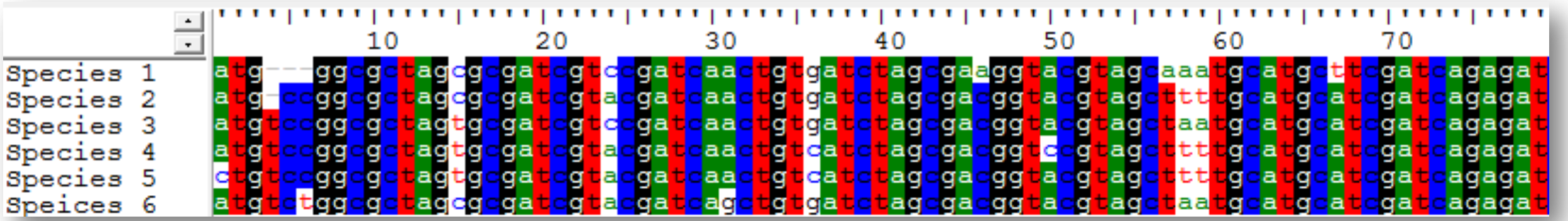
Discrimination (but not too much)

Fail: Sequence shows no conservation, impossible for PCR, but good as barcode



Discrimination (but not too much)

Win: Sequence shows some (ideally ~70%) conservation, good for PCR, good as barcode



Robustness

Since barcoding protocols (typically) amplify a region of DNA by PCR, also need to select a locus that amplifies reliably and sequences well

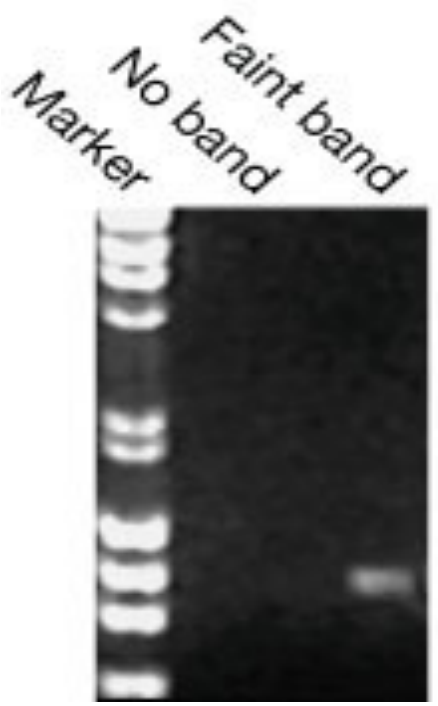


Photo credit
https://www.bio-rad.com/es-mx/applications-technologies/pcr-troubleshooting?ID=LUSO3HC4S

Choosing a barcoding locus

Cytochrome Oxidase C subunit 1 (COI):

- Universal
- Discriminating
- Robust

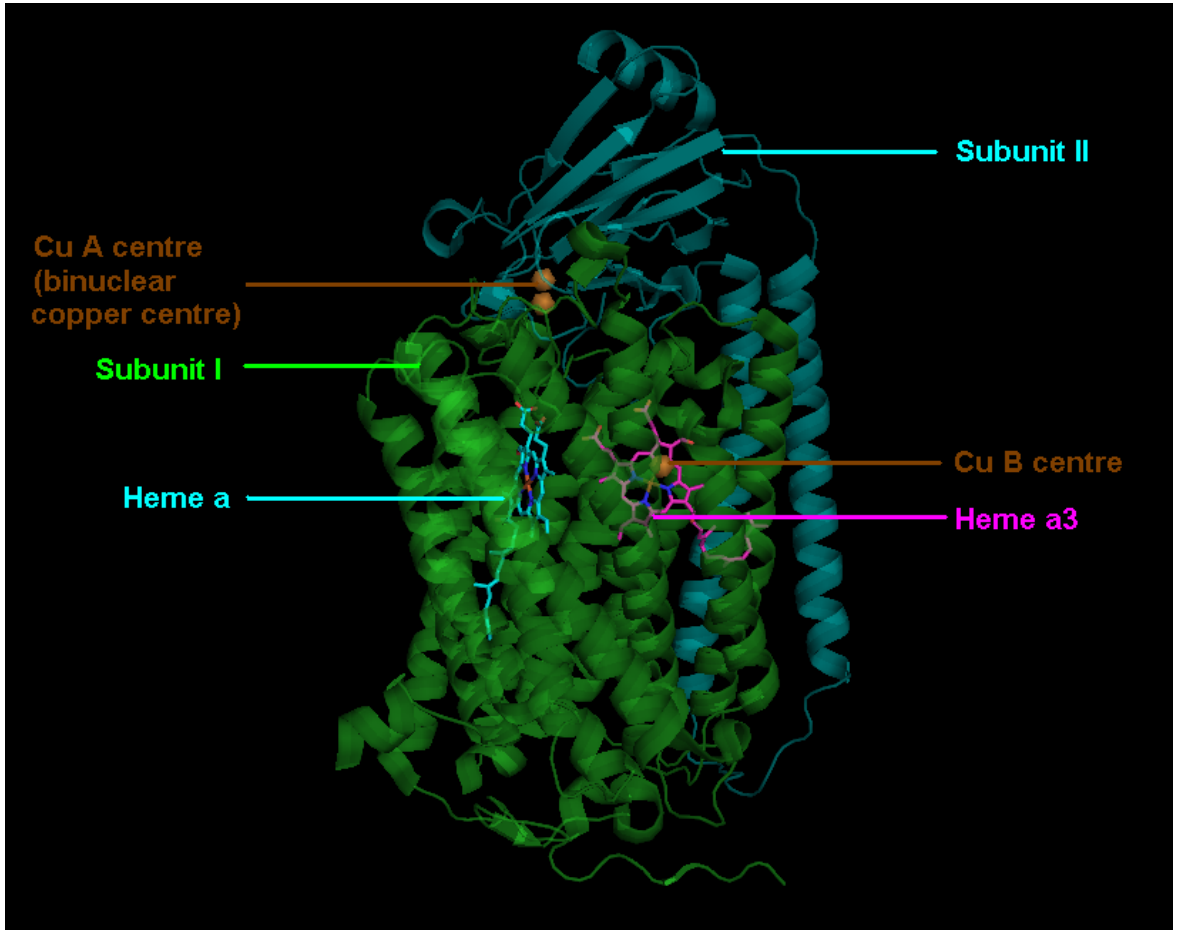
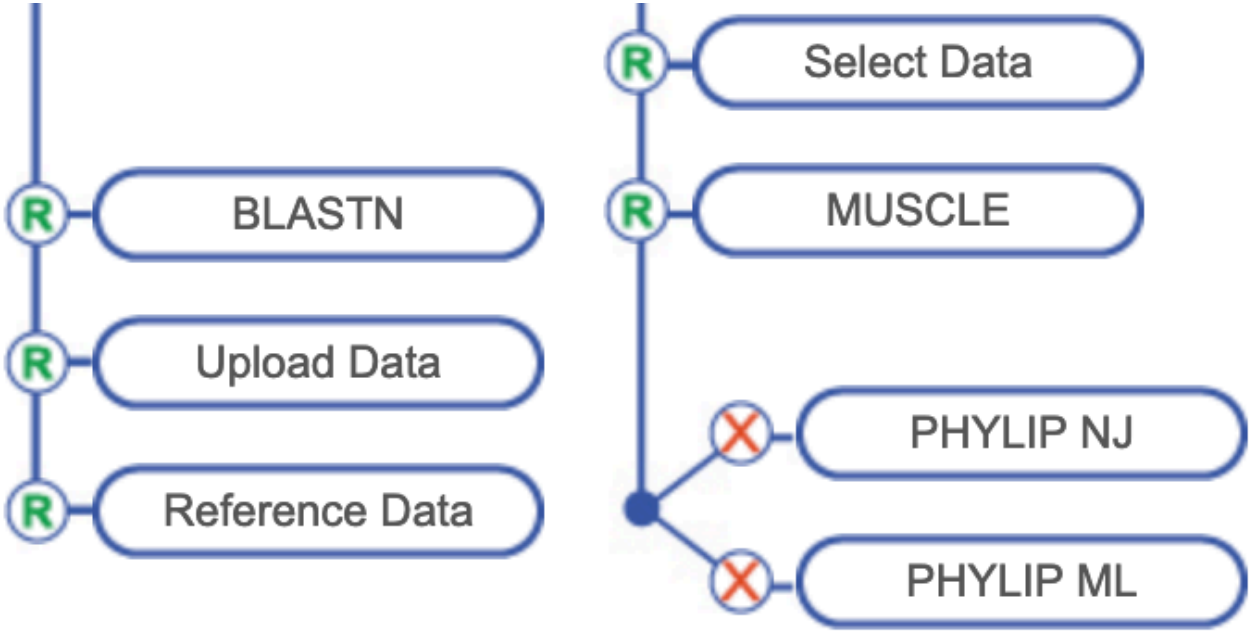


Photo credit:
https://en.wikipedia.org/wiki/Cytochrome_c_oxidase#/media/File:Cmplx4.PNG

Reference data and phylogenetics



Experimental components/design

Materials

- We have DNA from unknown mosquito samples
- We can obtain DNA from known samples

Hypothesis

- We can use computational methods (BLAST/phylogenetic analysis) to infer the species

Controls

- We have sensitivity controls (sequence quality, BLAST parameters)
- We have outgroup sequences (non-mosquito, negative controls) and known samples (positive controls)

Intro to Multiple Sequence Alignment

Multiple sequence alignment

To compare sequence features (nucleotides, amino acids, etc.) we need to line them up

Photo credit:
https://www.pexels.com/photo/jigsaw-puzzle-1586950/

Warning: Analogy

(useful for discussion but not the whole picture)

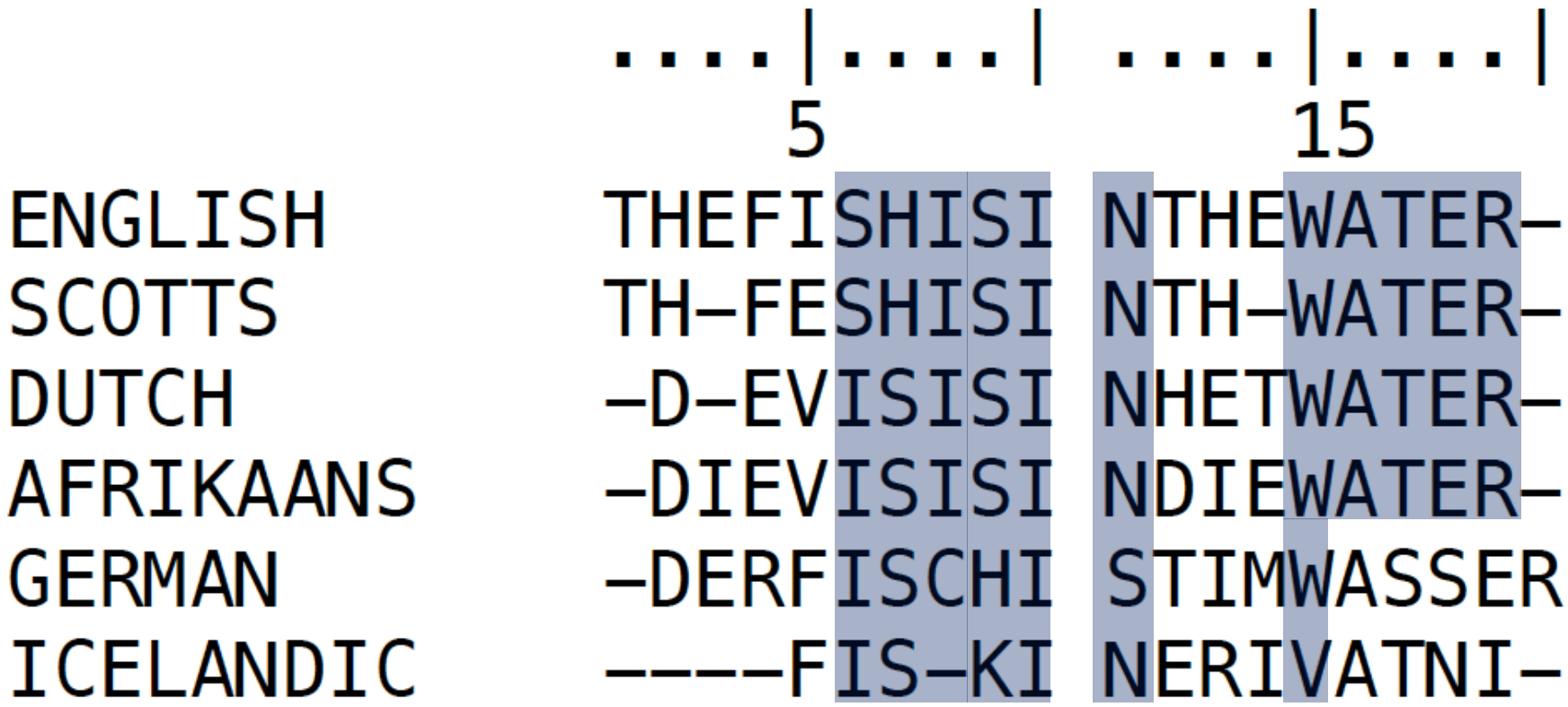
Multiple sequence alignment

THEFISHISINTHEWATER
TH’FESHISINTH’WATER
DEVISISINHETWATER
DIEVISISINDIEWATER
DERFISCHISTIMWASSER
FISKINERÍVATNI

Multiple sequence alignment

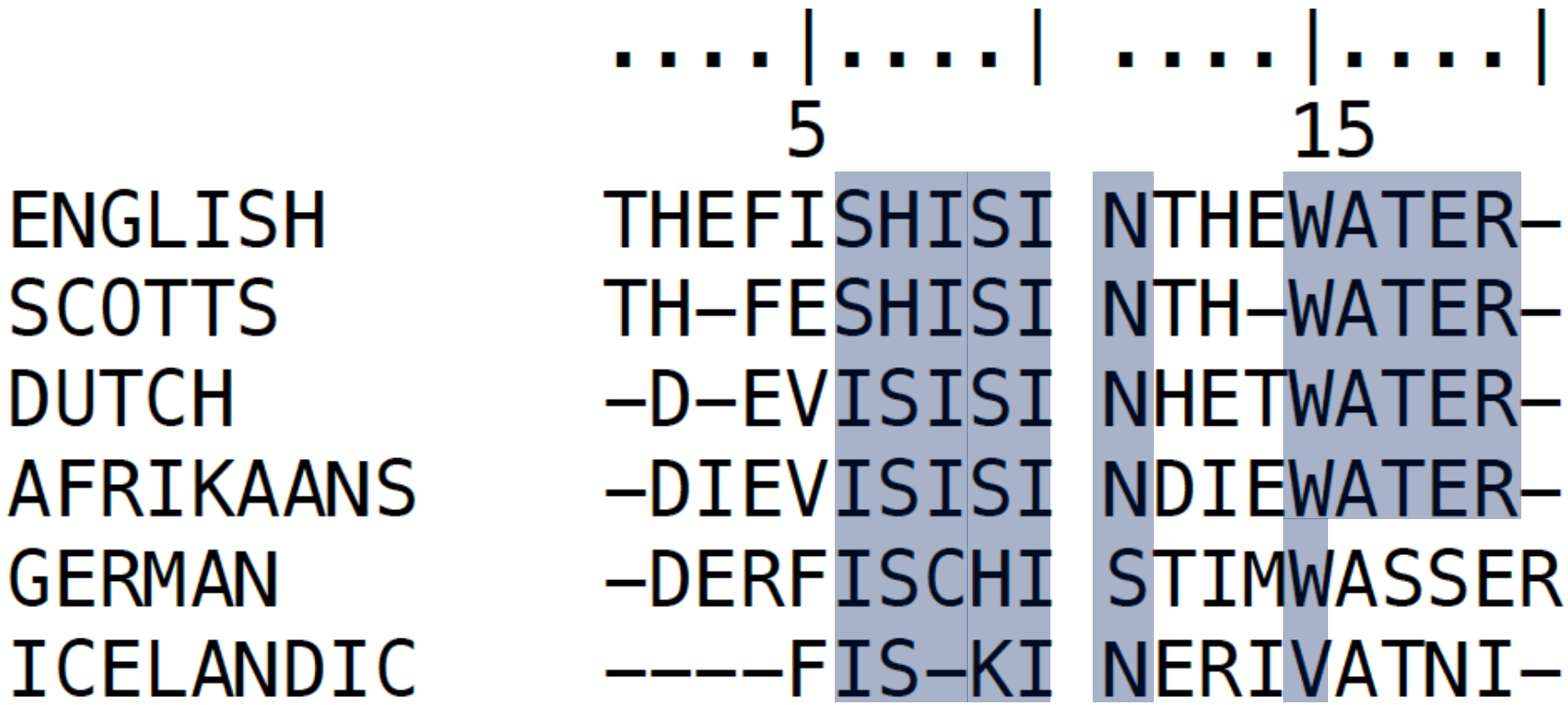
```
                  ....|....|  ....|....|
                      5           15
ENGLISH       THEFISHISI NTHEWATER-
SCOTTS        TH-FESHISI NTH-WATER-
DUTCH         -D-EVISISI NHETWATER-
AFRIKAANS     -DIEVISISI NDIEWATER-
GERMAN        -DERFISCHI STIMWASSER
ICELANDIC     ----FIS-KI NERIVATNI-
```

Multiple sequence alignment



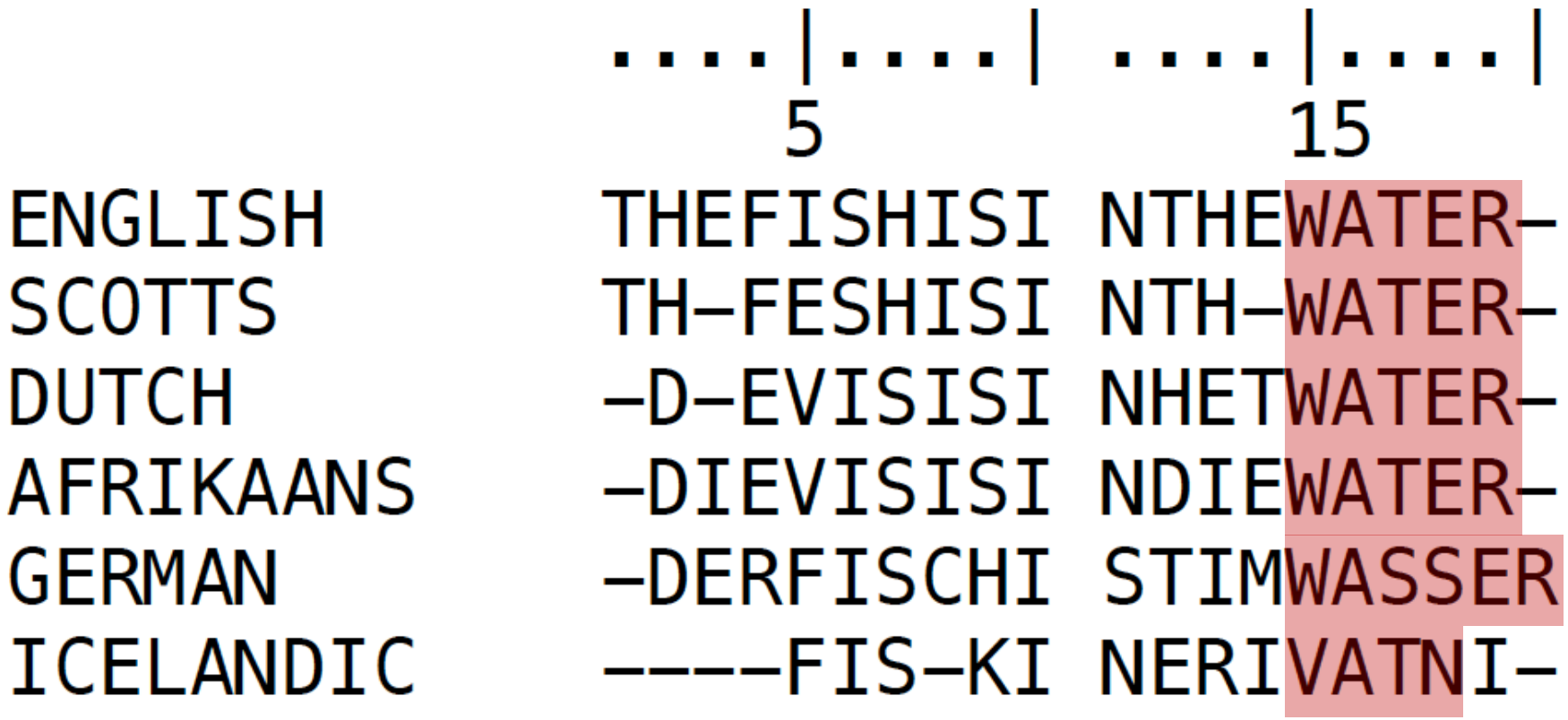
Colored at 60% identity

Multiple sequence alignment



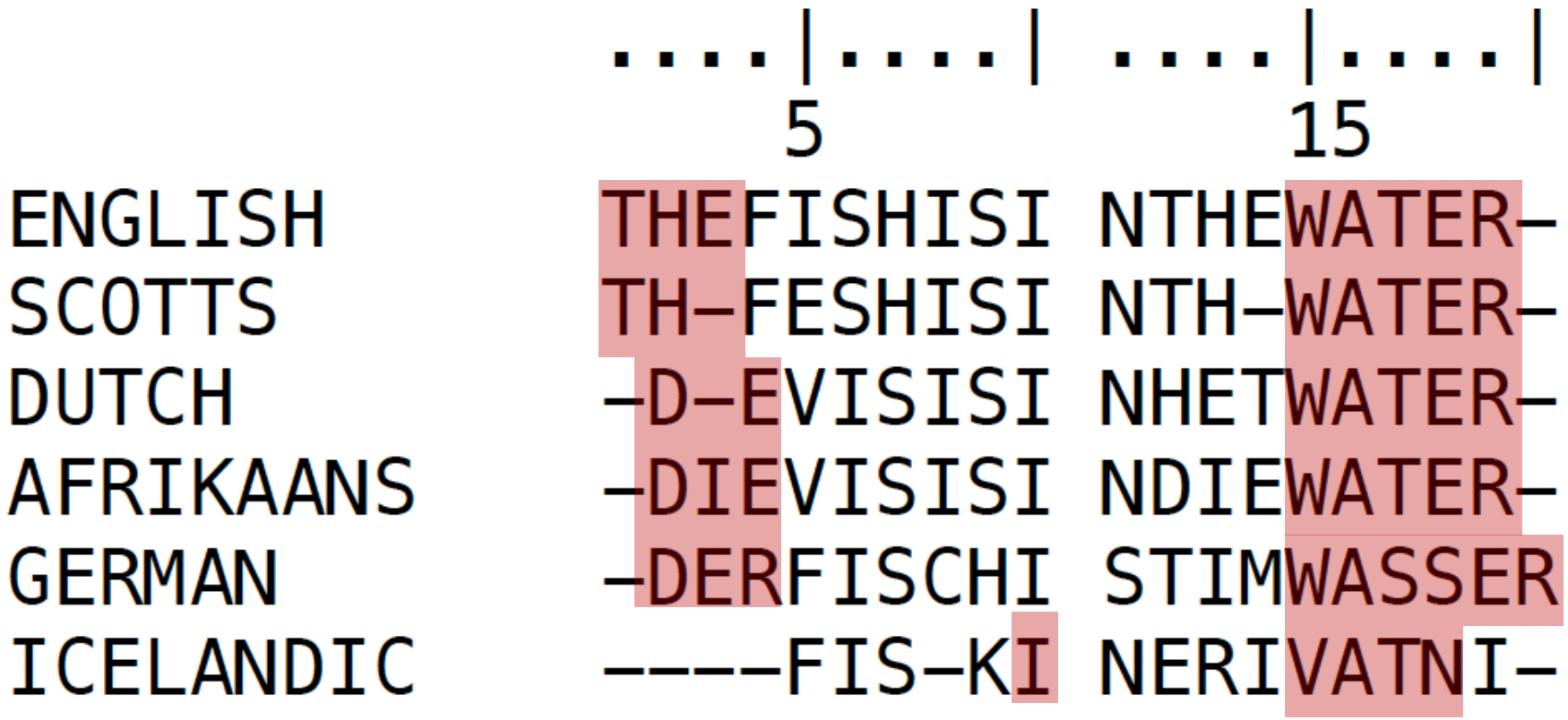
There is more information here …

Multiple sequence alignment



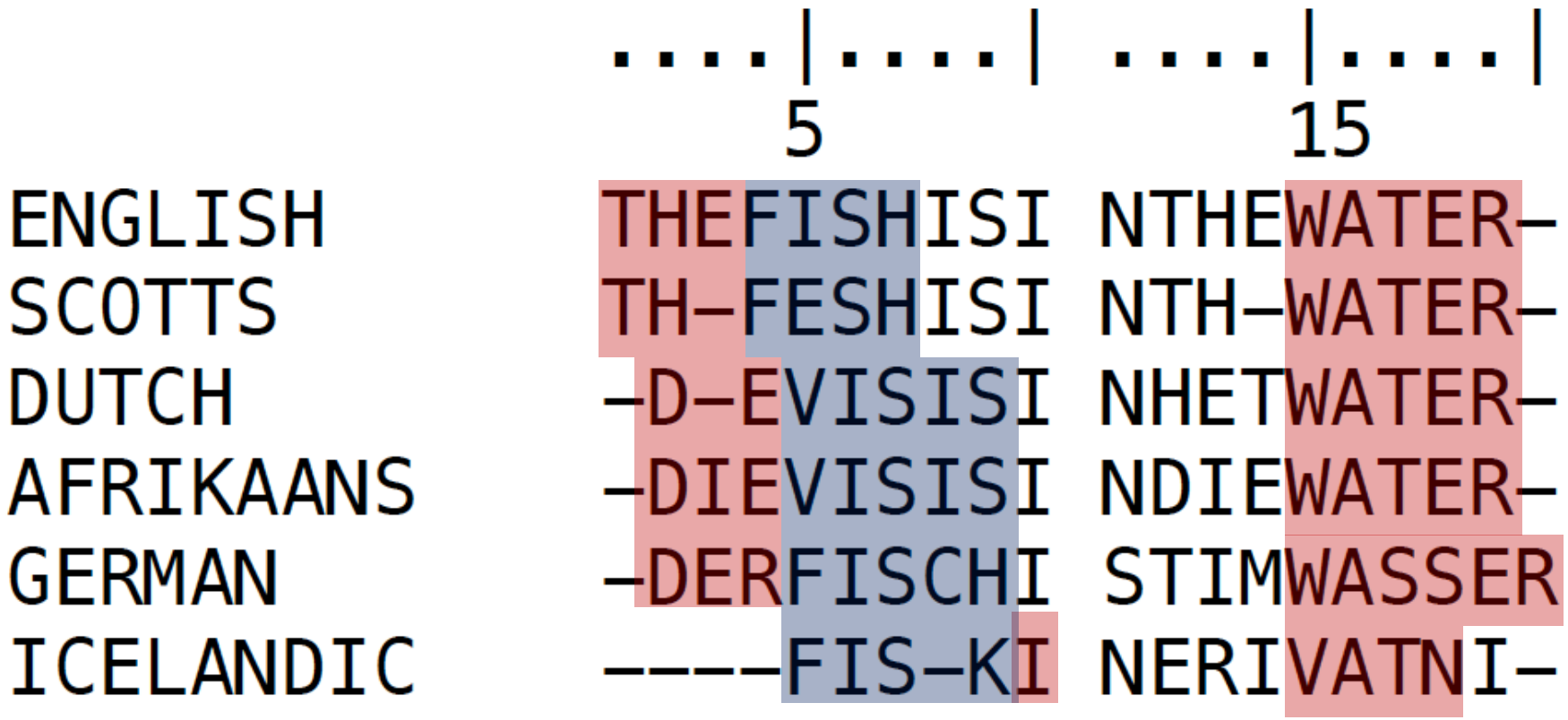
There is more information here …

Multiple sequence alignment



There is more information here …

Multiple sequence alignment



There is more information here …

Multiple sequence alignment

Number of (global) alignments for 2 sequences of length n

$$\frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{\pi n}}$$

Photo credit:
https://www.pexels.com/photo/jigsaw-puzzle-1586950/

Multiple sequence alignment

Number of (global) alignments for 2 sequences of length n

$$\frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{\pi n}}$$

So, for 2 sequences (n=100) ≈ 10^{77} *

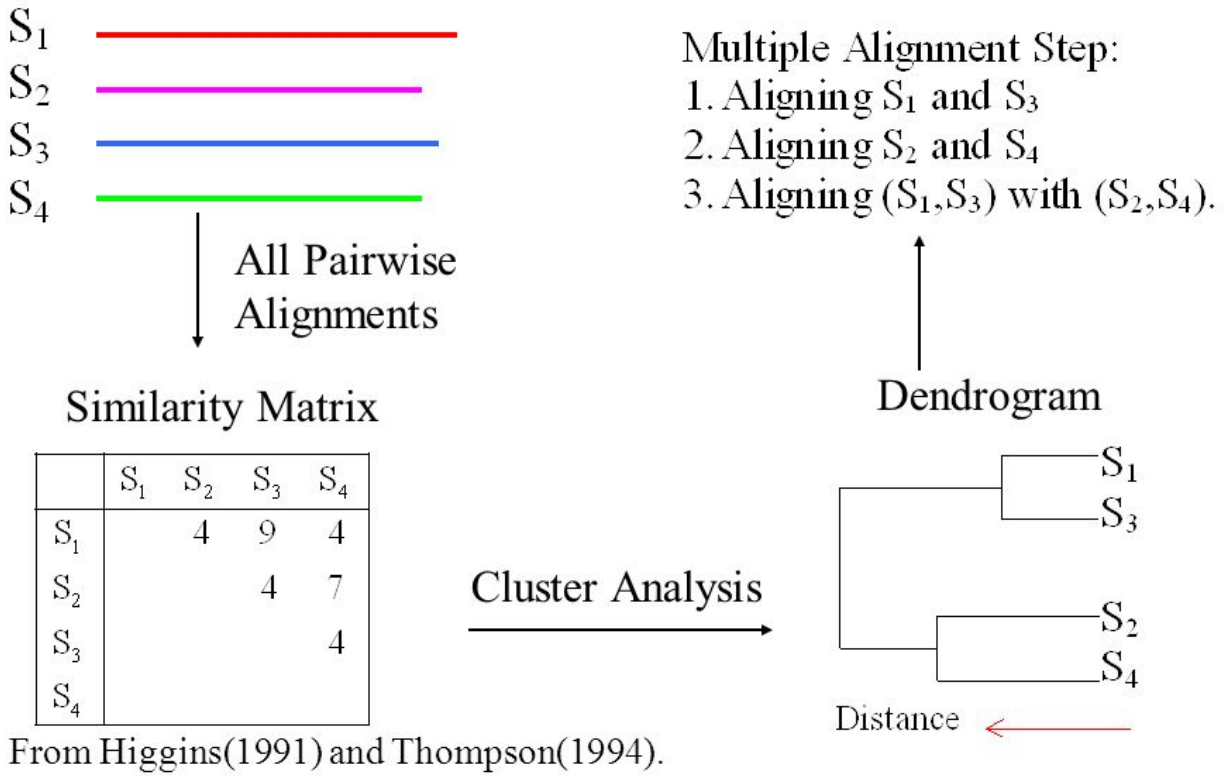
- We need software!

Photo credit:
https://www.pexels.com/photo/jigsaw-puzzle-1586950/

*(some are trivial)

CLUSTAL alignment algorithm

ClustalW steps

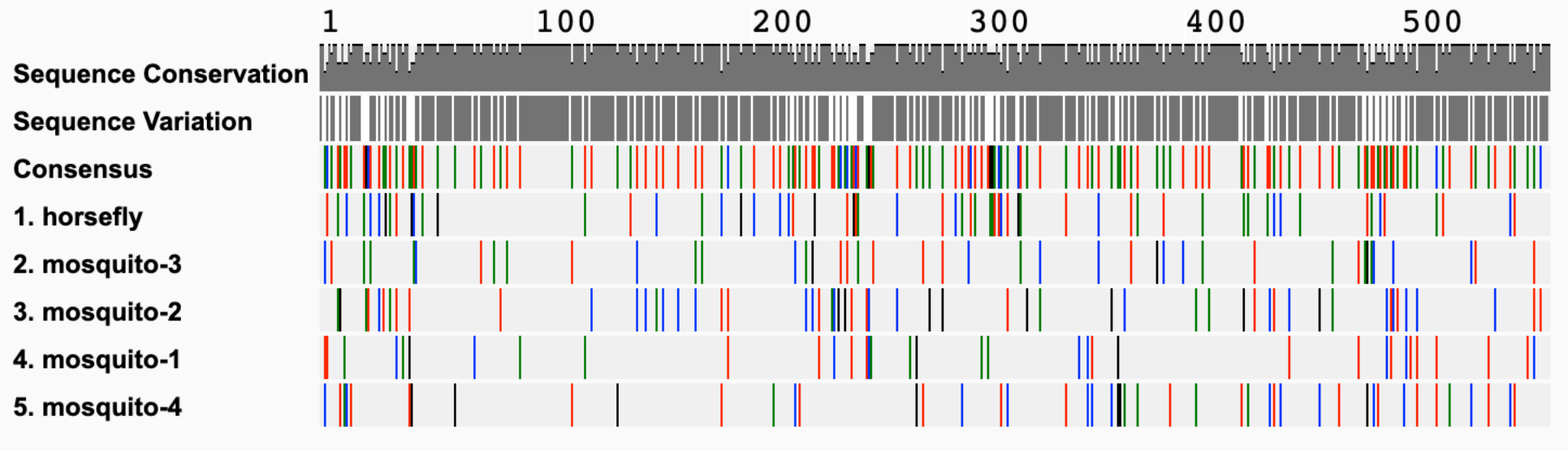


	S1	S2	S3	S4
S1		4	9	4
S2			4	7
S3				4
S4				

From Higgins(1991) and Thompson(1994).

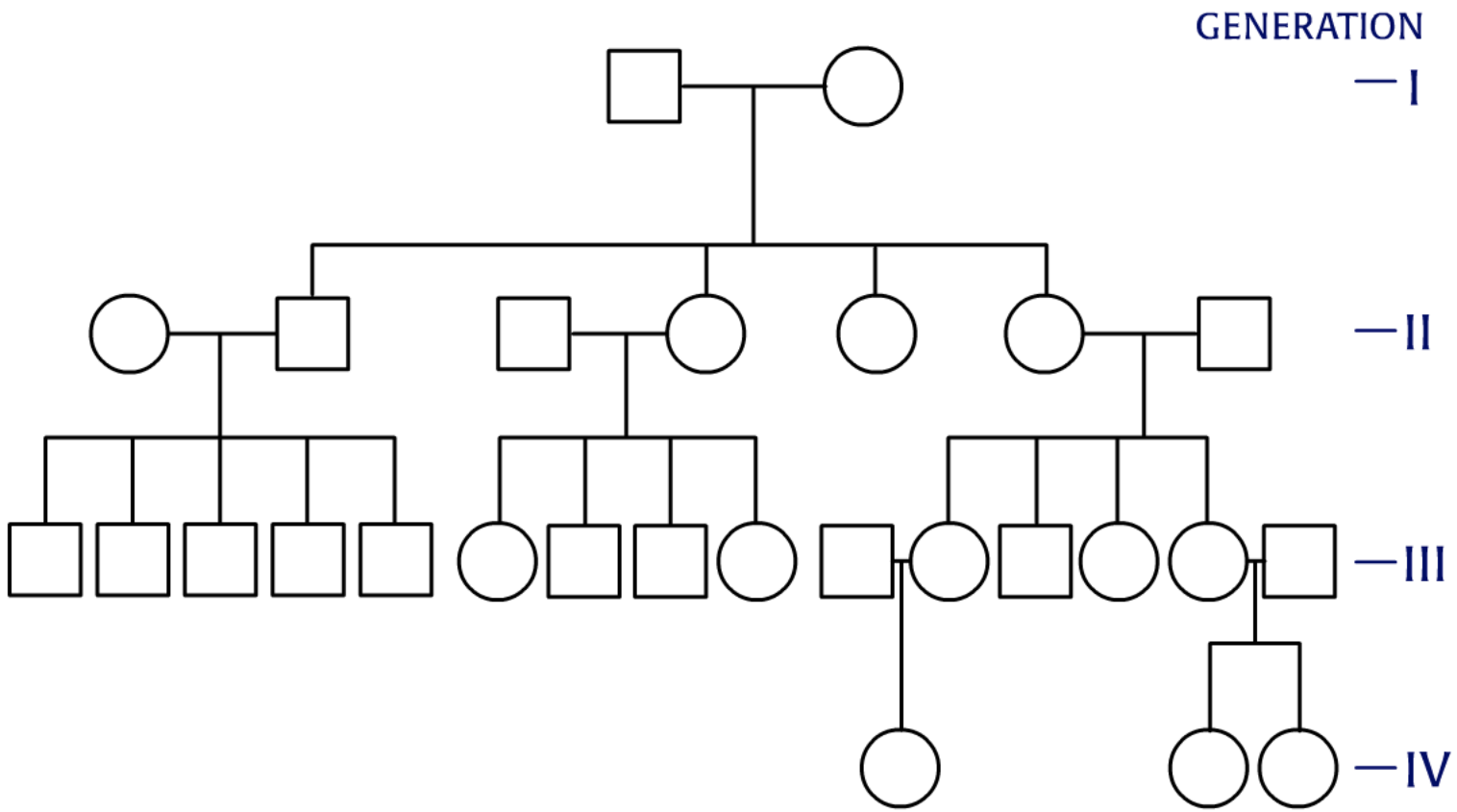
Photo credit:
https://slideplayer.com/slide/5155996/

Multiple sequence alignment



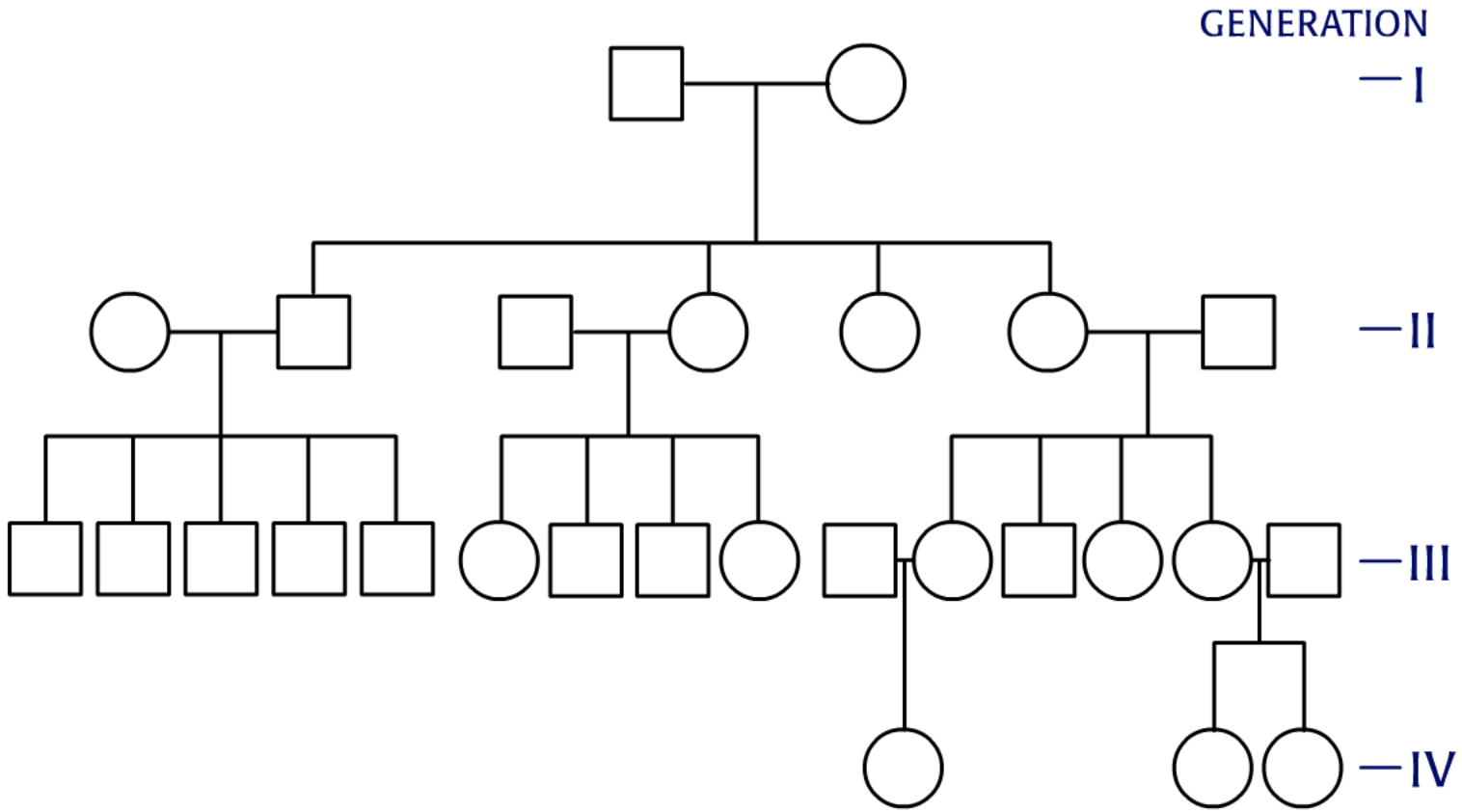
Making phylogenetic trees

Tree relationships



Tree relationships

This is not a phylogenetic tree



Phylogenetic trees

A phylogenetic tree is a hypothesis about how species may be related

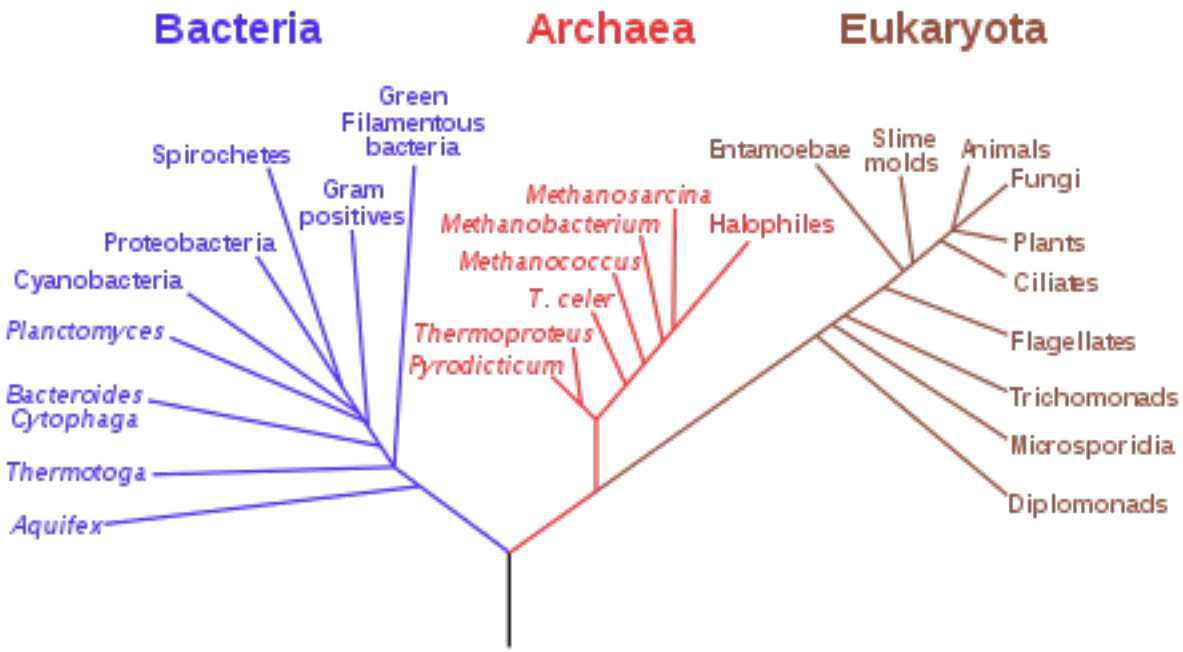


Photo credit
https://en.wikipedia.org/wiki/File:Phylogenetic_tree.svg

Phylogenetic trees

Trees can be created from (and therefore reflect) characteristics/traits

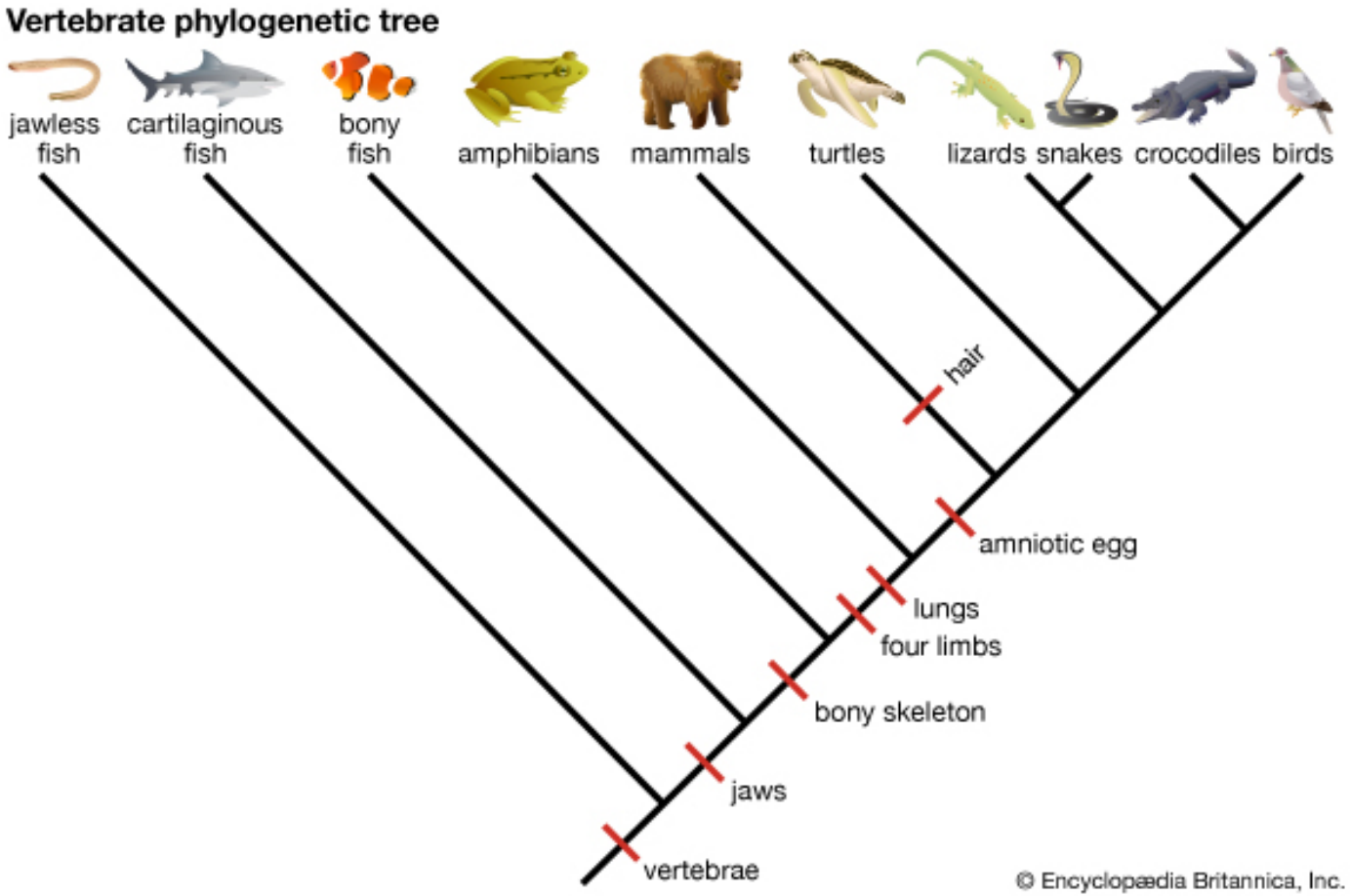


Photo credit
https://kids.britannica.com/students/assembly/view/235364

Phylogenetic trees

Trees can be created from (and therefore reflect) characteristics/traits

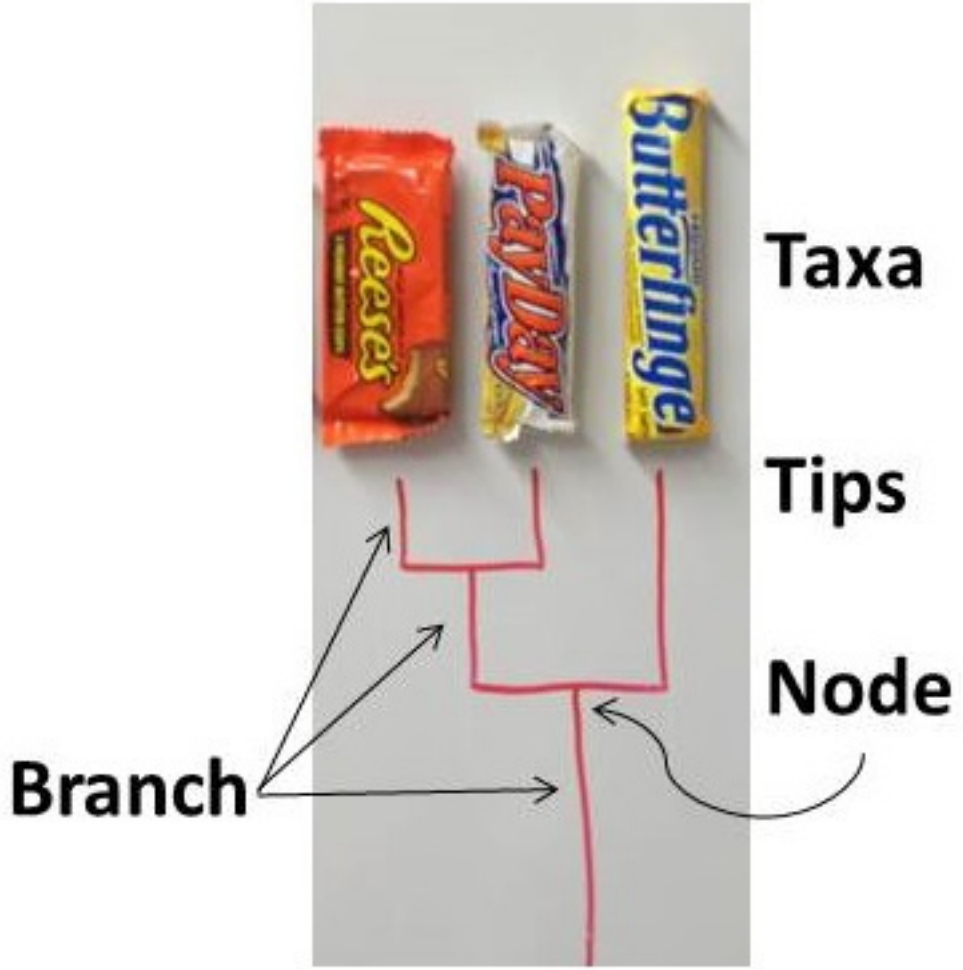


Photo credit
https://wildlifesnpits.wordpress.com/2014/03/22/understanding-phylogenies-terminology/

Phylogenetic trees

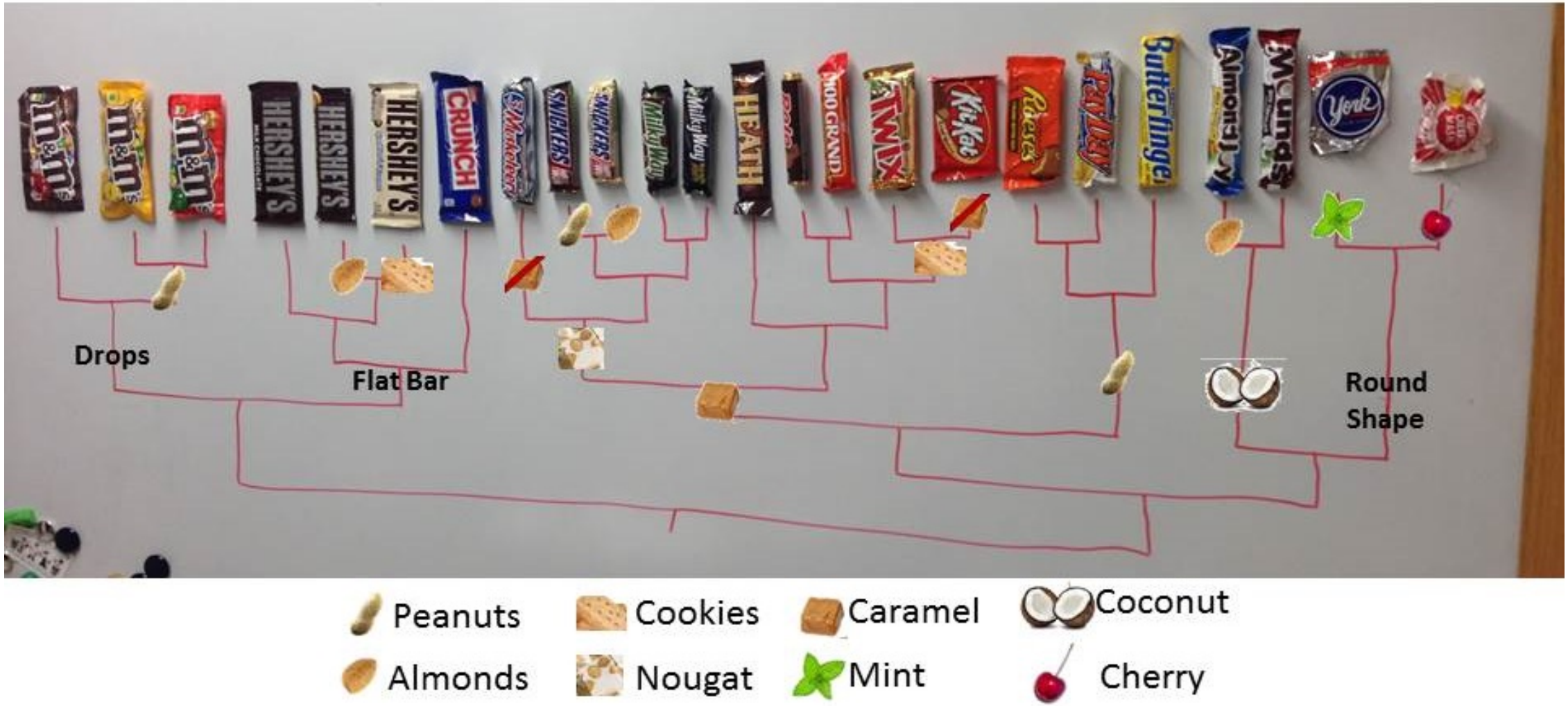


Photo credit
https://wildlifesnpits.wordpress.com/2014/03/22/understanding-phylogenies-terminology/

Phylogenetic trees

Tree Vocabulary

- Taxa: Individual species
- Tip: The endpoints of a tree
- Node: A point where branches split
- Branch: Any collection after a node

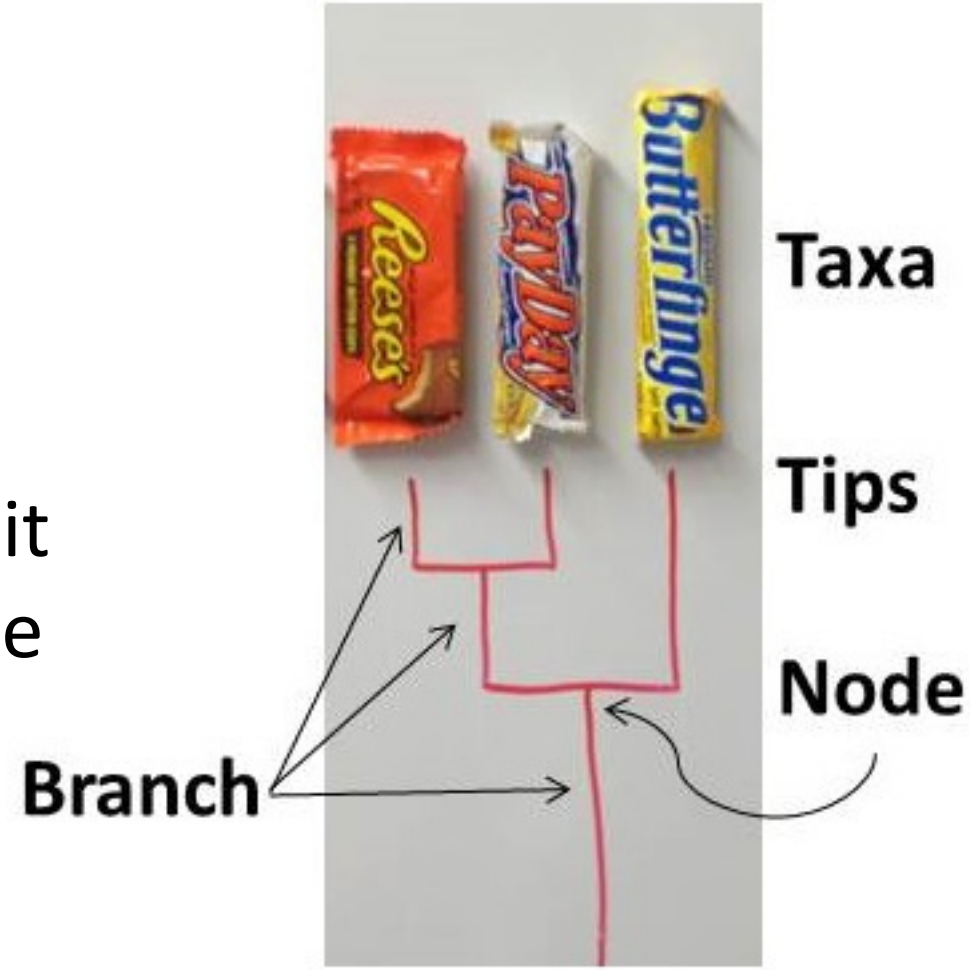


Photo credit
https://wildlifesnpits.wordpress.com/2014/03/22/understanding-phylogenies-terminology/

Tree experiments

- Experiment 1: Unknown Mosquitos and outgroup
  - Method: neighbor joining
- Experiment 2: Unknown Mosquitos and outgroup
  - Method: maximum likelihood
- Experiment 3: Unknown Mosquitos + BLAST hits and outgroup
  - Method: neighbor joining
- Experiment 4: Unknown Mosquitos + BLAST hits and outgroup
  - Method: maximum likelihood

Trees are also computationally intensive

Number of possible rooted trees for n sequences

$$= (2n-3)! / (2^{n-2} (n-2))!$$

2 sequences:	1
3 sequences:	3
4 sequences:	15
5 sequences:	105
6 sequences:	954
7 sequences:	10395
8 sequences:	135135
9 sequences:	2027025
10 sequences:	34459425
51 sequences:	$>10^{80}$ (nb of particles in the universe)

Photo credit:
https://www.slideshare.net/sebastiendelandtsheer/phylogenetics1

Tree building methods

- There is no one “best” method

Tree building methods

- There is no one “best” method

Neighbor joining:

- “distance-based” method of tree building

Tree building methods

- There is no one “best” method

Neighbor joining:

- “distance-based” method of tree building
- a matrix of the sequences is created based on distance between each pair of sequences in multiple alignment

Tree building methods

- There is no one “best” method

Neighbor joining:

- “distance-based” method of tree building
- a matrix of the sequences is created based on distance between each pair of sequences in multiple alignment
- distance is related to the number of mismatches between the sequences

Tree building methods

Bootstrap values (how we evaluate NJ trees):

- columns in the sequence alignment randomly resampled to make many new alignments (DNA subway does this 100x)

Tree building methods

Bootstrap values (how we evaluate NJ trees):

- columns in the sequence alignment randomly resampled to make many new alignments (DNA subway does this 100x)
- a matrix of the sequences is created

Tree building methods

Bootstrap values (how we evaluate NJ trees):

- columns in the sequence alignment randomly resampled to make many new alignments (DNA subway does this 100x)
- a matrix of the sequences is created
- each bootstrap value is the number of times that particular relationship appears in the 100 resampled trees

Tree building methods

Bootstrap values (how we evaluate NJ trees):

- columns in the sequence alignment randomly resampled to make many new alignments (DNA subway does this 100x)
- a matrix of the sequences is created
- each bootstrap value is the number of times that particular relationship appears in the 100 resampled trees
- Values of 70 and above are “plausible”; above 95 considered highly supported

Tree building methods

Maximum likelihood:

- Attempts to take into account observed patterns of how nucleotides and amino acids change over time. For instance, mutations from C to T are more common than mutations of C to A.

Tree building methods

Maximum likelihood:

- Attempts to take into account observed patterns of how nucleotides and amino acids change over time. For instance, mutations from C to T are more common than mutations of C to A.
- The tree with highest overall likelihood score is accepted as the best estimate of the relationships between sequences.

Tree building methods

Maximum likelihood:

- Attempts to take into account observed patterns of how nucleotides and amino acids change over time. For instance, mutations from C to T are more common than mutations of C to A.
- The tree with highest overall likelihood score is accepted as the best estimate of the relationships between sequences.
- The length of the branches between nodes is also a measure of the confidence of the relationships. Very short branches separating sequences have lower confidence and the relationships are less certain, while longer branches are better supported.