Persistent Homology in Genomics

Plan: ① Review of Evolutionary Biology

② Appearance of non-trivial Topology & how Persistent Homology can help

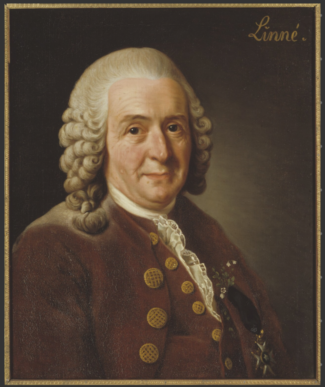
③ Example: Influenza

References: [1] "Topology of viral evolution" (Chan, Carlsson, Rabadán)

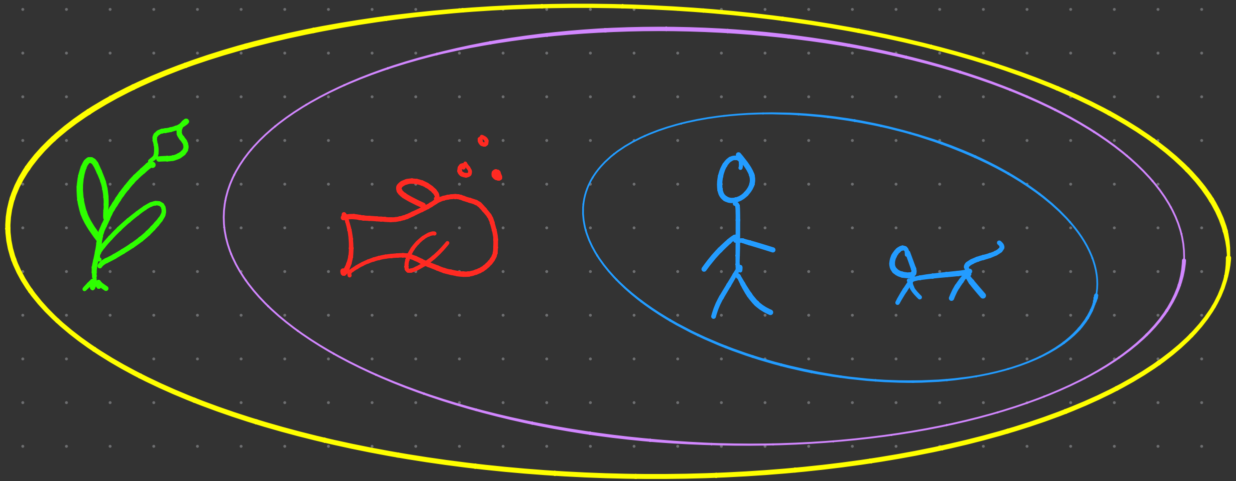
[2] "TDA for Genomics and Evolution" (Rabadán, Blumberg)

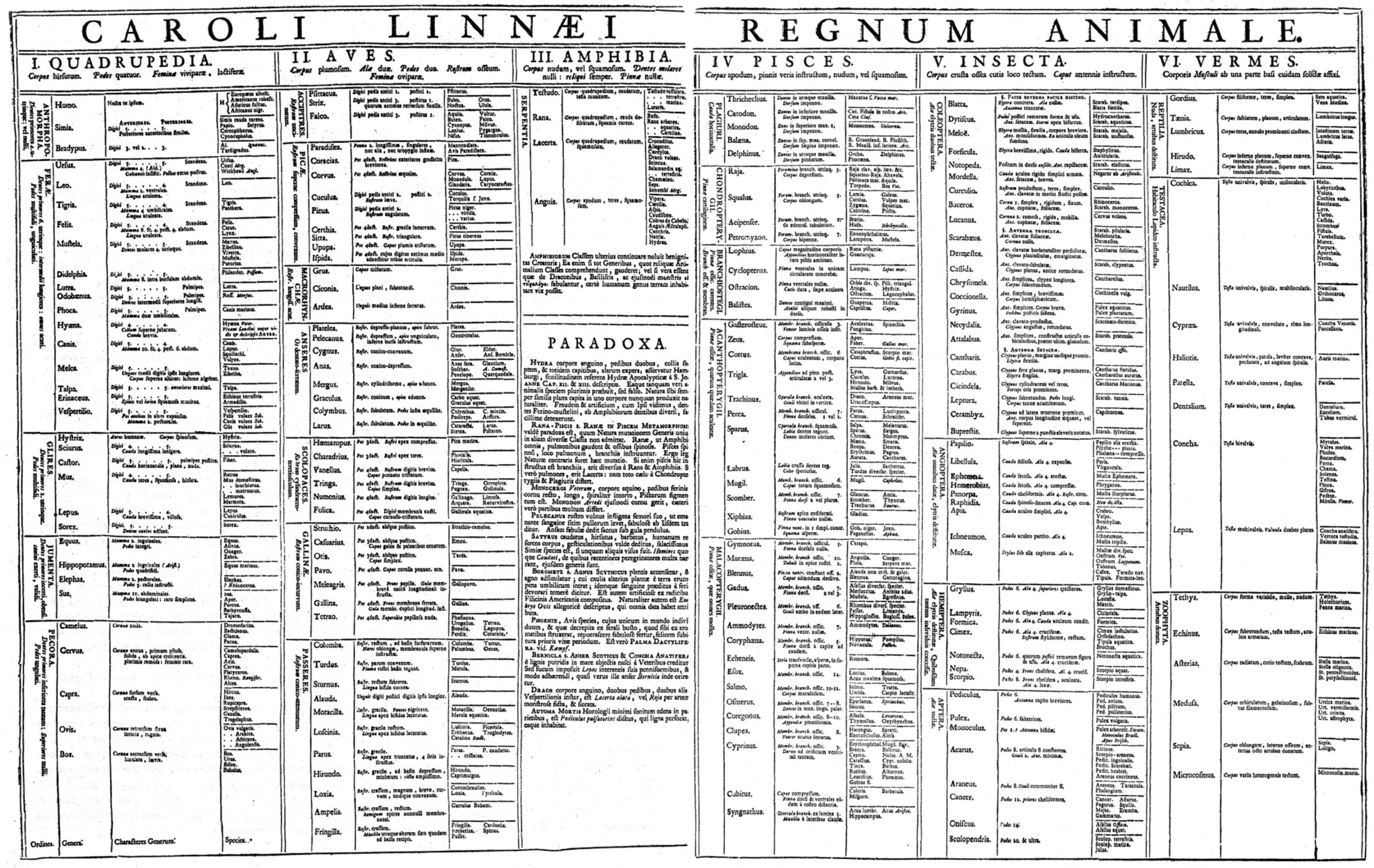
① Review of Evolutionary Biology

Carl von Linné 1735: Attempt to classify all living organisms on earth



Source: Wiki





CAROLI LINNÆI REGNUM ANIMALE.

I. QUADRUPEDIA.

Corpus hirſutum. *Pedes* quatuor. *Feminæ* viviparæ, laſtiferæ.

- ANTHROPOMORPHA: Homo. Simia. Bradypus.
- FERÆ: Urſus. Leo. Tigris. Felis. Muſtela. Didelphis. Lutra. Odobænus. Phoca. Hyæna. Canis. Meles. Talpa. Erinaceus. Veſpertilio.
- GLIRES: Hyſtrix. Sciurus. Caſtor. Mus. Lepus. Sorex.
- JUMENTA: Equus. Hippopotamus. Elephas. Sus.
- PECORA: Camelus. Cervus. Capra. Ovis. Bos.

II. AVES.

Corpus plumoſum. *Alæ* duæ. *Pedes* duo. *Roſtrum* oſſeum. *Feminæ* oviparæ.

- ACCIPITRES: Pſittacus. Strix. Falco.
- PICÆ: Paradiſæa. Coracias. Corvus. Cuculus. Picus. Certhia. Sitta. Upupa. Iſpida.
- MACRORHYNCHÆ: Grus. Ciconia. Ardea.
- ANSERES: Platelea. Pelecanus. Cygnus. Anas. Mergus. Graculus. Colymbus. Larus.
- SCOLOPACES: Hæmatopus. Charadrius. Vanellus. Tringa. Numenius. Fulica.
- GALLINÆ: Struthio. Caſuarius. Otis. Pavo. Meleagris. Gallina. Tetrao.
- PASSERES: Columba. Turdus. Sturnus. Alauda. Motacilla. Luſcinia. Parus. Hirundo. Loxia. Ampelis. Fringilla.

III. AMPHIBIA.

Corpus nudum, vel ſquamoſum. *Dentes molares* nulli: *reliqui* ſemper. *Pinnæ* nullæ.

- SERPENTIA: Teſtudo. Rana. Lacerta. Anguis.

AMPHIBIORUM Claſſem ulterius continuare noluit benignitas Creatoris; Ea enim ſi tot Generibus, quot reliquæ Animalium Claſſes comprehendunt, gauderet; vel ſi vera eſſent quæ de Draconibus, Baſiliſcis, ac ejuſmodi monſtris a τεραλόγοι fabulantur, certè humanum genus terram inhabitare vix poſſet.

PARADOXA.

HYDRA corpore anguino, pedibus duobus, collis ſeptem, & totidem capitibus, alarum expers, aſſervatur Hamburgi, ſimilitudinem referens Hydræ Apocalypticæ à S. JOANNE CAP. XII. & XIII. deſcriptæ. Eaque tanquam veri animalis ſpeciem plurimis præbuit, ſed falſo. Natura ſibi ſemper ſimilis plura capita in uno corpore nunquam produxit naturaliter. Fraudem & artificium, cum Ipſi vidimus, dentes Ferino-muſtelini, ab Amphibiorum dentibus diverſi, facillime detexerunt.

RANA-PISCIS s. RANÆ IN PISCEM METAMORPHOSIS valdè paradoxa eſt, quum Natura mutationem Generis unius in aliam diverſæ Claſſis non admittat. Ranæ, ut Amphibia omnia, pulmonibus gaudent & oſſibus ſpinoſis. Piſces ſpinoſi, loco pulmonum, branchiis inſtruuntur. Ergo leg Naturæ contraria foret hæc mutatio. Si enim piſcis hic inſtructus eſt branchiis, erit diverſus à Rana & Amphibiis. S verò pulmones, erit Lacerta: nam toto cælo à Chondropterygiis & Plagiuris differt.

MONOCEROS *Veterum*, corpore equino, pedibus ferinis cornu recto, longo, ſpiraliter intorto, Pictorum figmentum eſt. MONODON *Artedi* ejuſmodi cornu gerit, cæteri verò partibus multum differt.

PELECANUS roſtro vulnus infligens femori ſuo, ut emanante ſanguine ſitim pullorum levet, fabuloſè ab Iiſdem traditur. Anſam fabulæ dedit ſaccus ſub gula pendulus.

SATYRUS caudatus, hirſutus, barbatus, humanum referens corpus, geſticulationibus valde deditus, ſalaciſſimus Simiæ ſpecies eſt, ſi unquam aliquis viſus fuit. *Homines* quoque *Caudati*, de quibus recentiores peregrinatores multa narrant, ejuſdem generis ſunt.

BOROMETZ s. AGNUS SCYTHICUS plantis acceſetur, & agno adſimilatur; cui caulis alterius plantæ è terra erumpens umbilicum intrat; idemque ſanguine præditus à feri devorari temerè dicitur. Eſt autem artificioſè ex radicibu Filicinis Americanis compoſitus. Naturaliter autem eſt *Embryo Ovis* allegoricè deſcriptus, qui omnia data habet attributa.

PHOENIX, Avis ſpecies, cujus unicum in mundo individuum, & quæ decrepita ex ferali buſto, quod ſibi ex aromatibus ſtruxerat, repueraſcere fabuloſè fertur, felicem ſubitura prioris vitæ periodum. Eſt verò PALMA DACTYLIFERA. vid. *Kæmpf.*

BERNICLA s. ANSER SCOTICUS & CONCHA ANATIFERA è lignis putridis in mare abjectis naſci à Veteribus creditur Sed fucum impoſuit *Lepas* interaneis ſuis penniformibus, & modo adhærendi, quaſi verus ille anſer *Bernicla* inde oriretur.

DRACO corpore anguino, duobus pedibus, duabus alis Veſpertilionis inſtar, eſt *Lacerta alata*, vel *Raja* per artem monſtroſè ficta, & ſiccata.

AUTOMA MORTIS Horologii minimi ſonitum edens in parietibus, eſt *Pediculus pulſatorius* dictus, qui ligna perforat, eaque inhabitat.

IV. PISCES.

Corpus apodum, pinnis veris inſtructum, nudum, vel ſquamoſum.

- PLAGIURI: Trichechus. Catodon. Monodon. Balæna. Delphinus.
- CHONDROPTERYGII: Raja. Squalus. Acipenſer. Petromyzon.
- BRANCHIOSTEGI: Lophius. Cyclopterus. Oſtracion. Baliſtes.
- ACANTHOPTERYGII: Gaſteroſteus. Zeus. Cottus. Trigla. Trachinus. Perca. Sparus. Labrus. Mugil. Scomber. Xiphias. Gobius.
- MALACOPTERYGII: Gymnotus. Muræna. Blennus. Gadus. Pleuronectes. Ammodytes. Coryphæna. Echeneis. Eſox. Salmo. Oſmerus. Coregonus. Clupea. Cyprinus. Cobitis. Syngnathus.

V. INSECTA.

Corpus cruſta oſſea cutis loco tectum. *Caput* antennis inſtructum.

- COLEOPTERA: Blatta. Dytiſcus. Meloë. Forficula. Notopeda. Mordella. Curculio. Buceros. Lucanus. Scarabæus. Dermeſtes. Caſſida. Chryſomela. Coccionella. Gyrinus. Necydalis. Attalabus. Cantharis. Carabus. Cicindela. Leptura. Cerambyx. Bupreſtis.
- ANGIOPTERA: Papilio. Libellula. Ephemera. Hemerobius. Panorpa. Raphidia. Apis. Ichneumon. Muſca.
- HEMIPTERA: Gryllus. Lampyris. Formica. Cimex. Notonecta. Nepa. Scorpio.
- APTERA: Pediculus. Pulex. Monoculus. Acarus. Araneus. Cancer. Oniſcus. Scolopendria.

VI. VERMES.

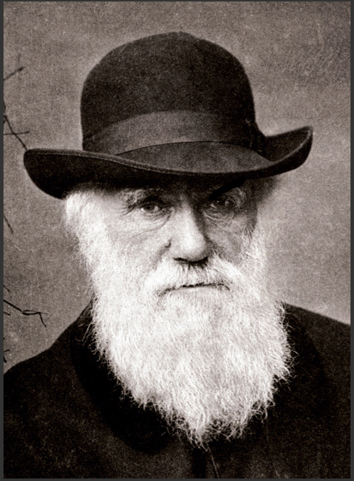
Corporis *Muſculi* ab una parte baſi cuidam ſolidæ affixi.

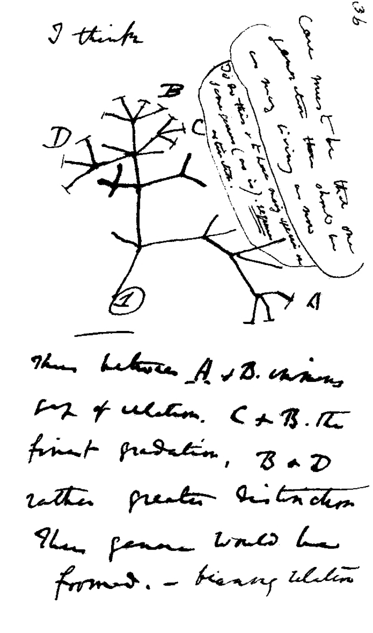
- REPTILIA: Gordius. Tænia. Lumbricus. Hirudo. Limax.
- TESTACEA: Cochlea. Nautilus. Cypræa. Haliotis. Patella. Dentalium. Concha. Lepas.
- ZOOPHYTA: Tethys. Echinus. Aſterias. Meduſa. Sepia. Microcoſmus.

Ordines. Genera. Characteres Generum. Species.

Source: Wiki

Charles Darwin 1859: This clustering also reflects the <u>origin</u> of these species.





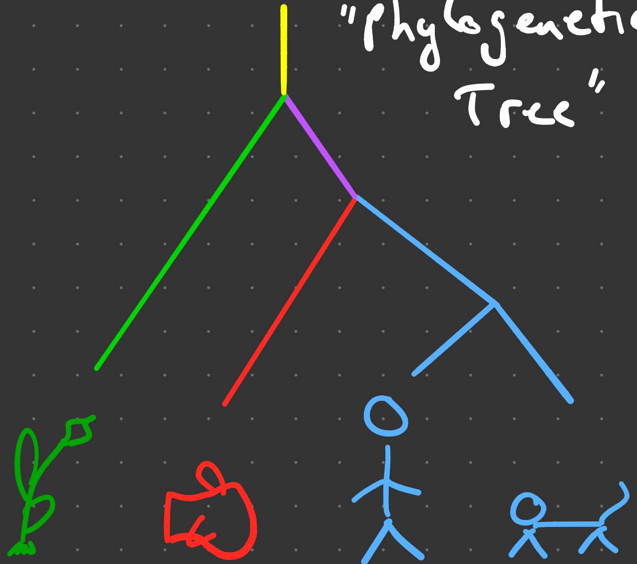
136

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.

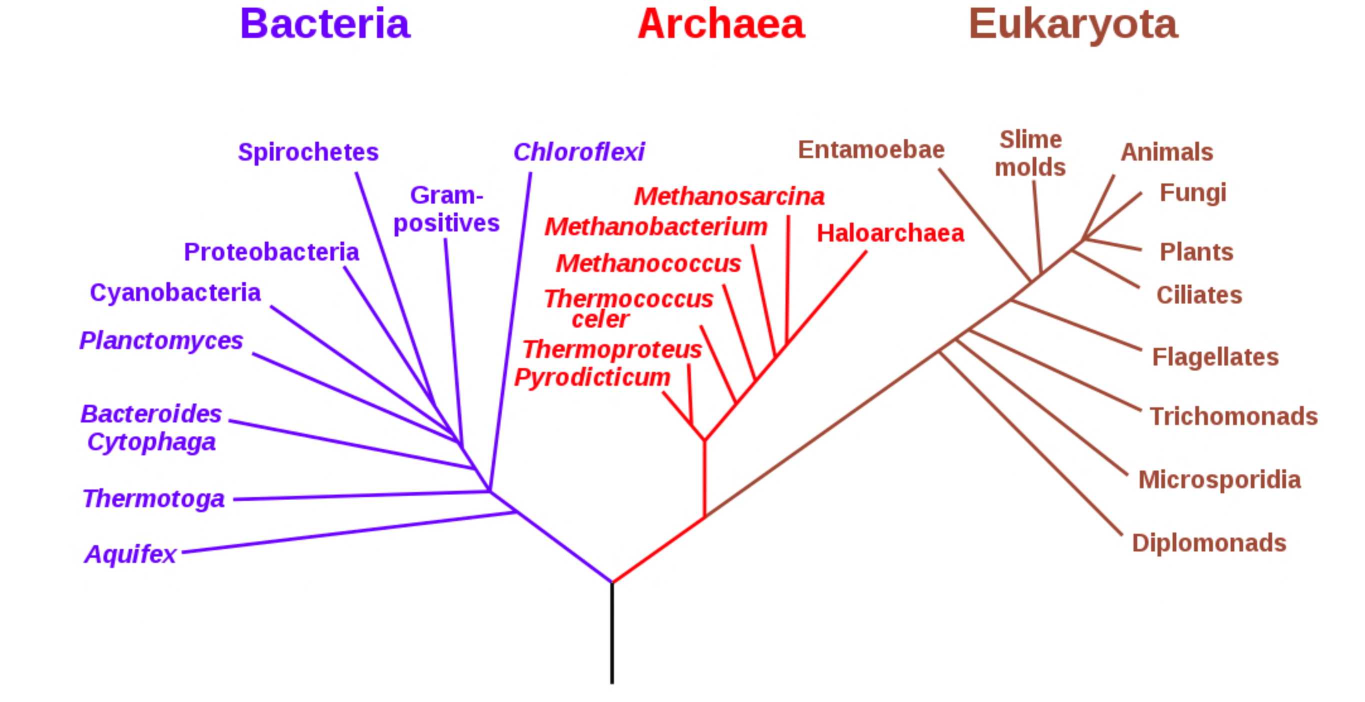
Thus between A & B. immense gap of relation. C & B. the finest gradation, B & D rather greater distinction Thus genera would be formed. — bearing relation

Source: Wiki



"Phylogenetic Tree"

Since the 1970s: Availability of molecular data for determining evolutionary history



Source: Wiki

How does this work?

All information about an organism is stored in a DNA or RNA molecule

DNA

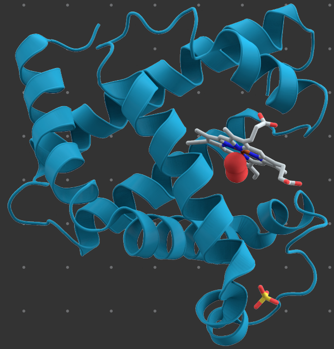
. . . A T G C T G C C A . . .
| | | | | | | | |
. . . T G C A T G T C A . . .

RNA . . . G U C A U A A G . . .

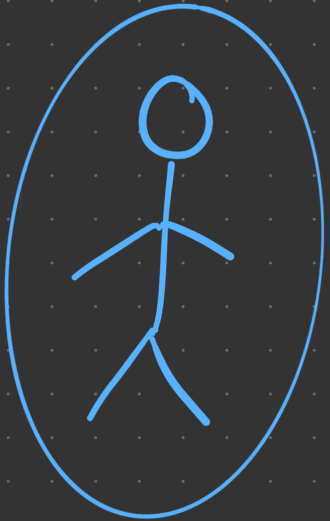
Adenin
Cytosin
Guanin
Thymin

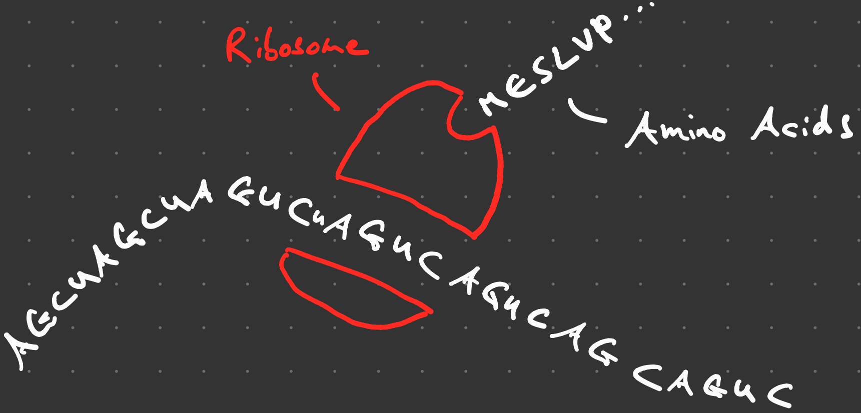
Uracil

Protein - Biosynthesis



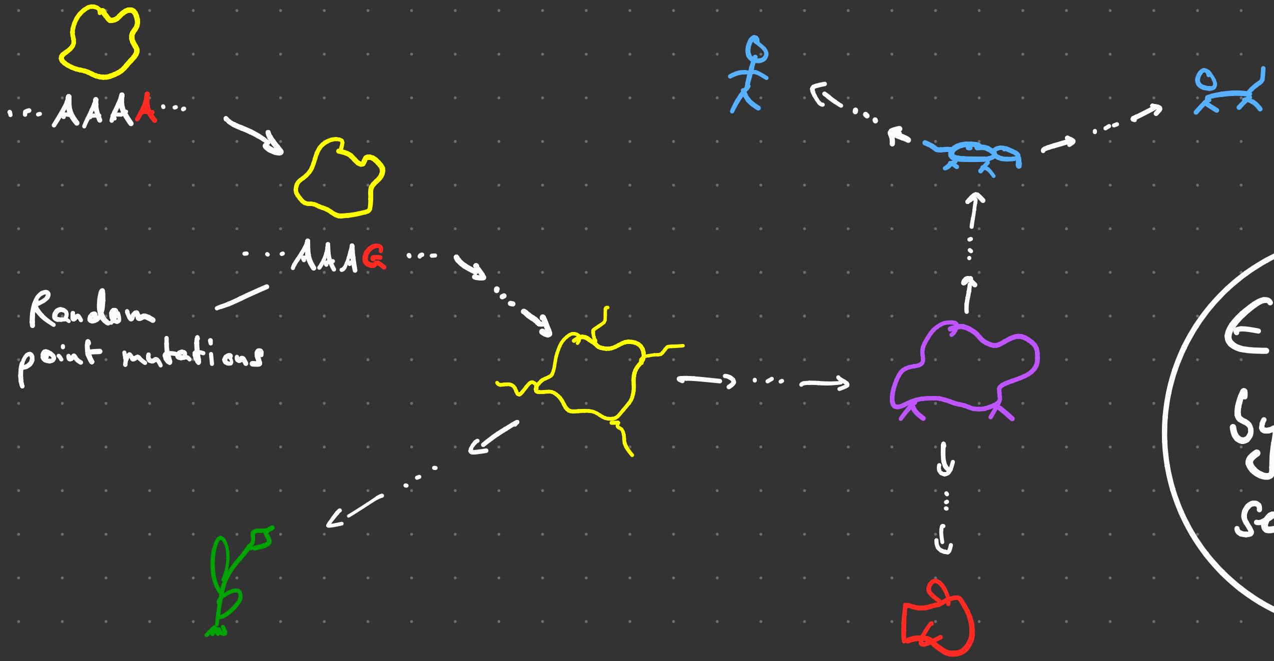
Source: Wiki





Simplest mode of reproduction : Clones

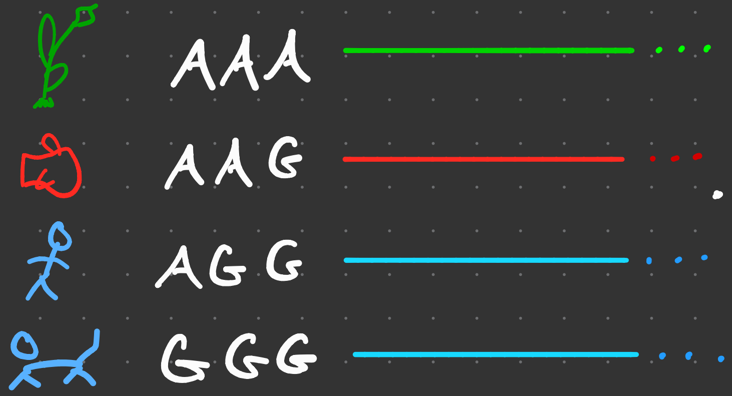
(Not a realistic depiction !)



Evolution by natural selection

Conversely, given

- A number of species
- together with their genomes AAA, AAG, AGG, GGG
- and a way to <u>compare them</u>



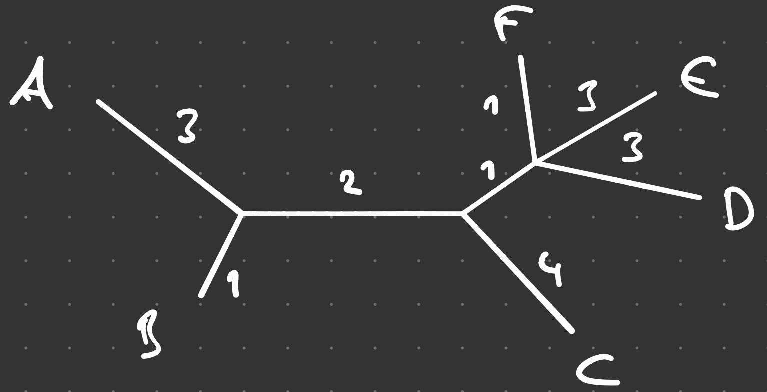
What is the tree?

Comparison Given an alphabet $\mathcal{A}$ (here $\#\mathcal{A} = 4$) we can define the **Hamming-distance** on $\mathcal{A}^n$, $n \in \mathbb{N}$, as

$$\begin{aligned} d: \mathcal{A}^n \times \mathcal{A}^n &\longrightarrow \mathbb{R} \qquad \left(\begin{matrix} x = AAA \\ y = AAG \end{matrix} \rightarrow d(x,y) = 1 \right) \\ (x, y) &\longmapsto d(x,y) \\ &= \#\{ i \in \{1, \dots, n\} \mid x_i \neq y_i \} \end{aligned}$$

$\Longrightarrow$ For a number of genomes of given length, we get a finite metric space (X, d_X).

Def. A tree is a finite, weighted and connected graph without loops and s.t. all vertices have either degree 1 or ≥ 3.



⟹ We get a finite set $\{A, \dots, F\}$ with a "tree metric"

Question: Is there an isometry (at least approximately)

$$(X, d_{Ham}) \xrightarrow{\sim} (X, d_{Tree}) \quad ?$$

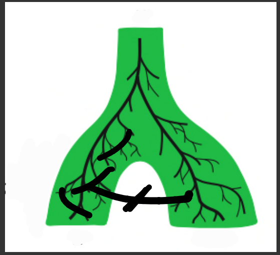
Facts:

- Not every metric arises from a tree metric
- There are (heuristic ?) algorithms to construct the tree, if possible

→ "Neighbor-Joining"

$$|d_{Ham}(x,y) - d_{Tree}(x,y)| \leq \frac{1}{2} \min d_{Tree}(x,y)$$

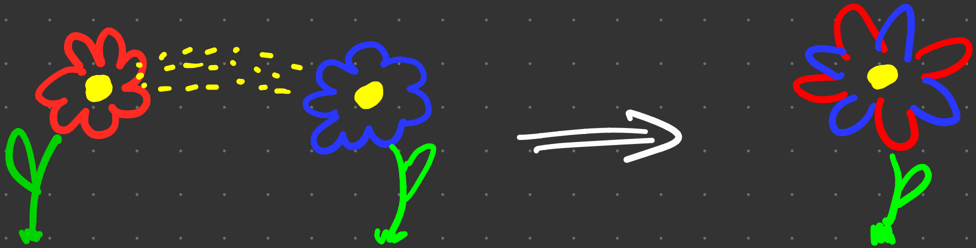
② Appearance of Topology



Source: [2]

- Implication of the "tree-paradigm" is that species have to be reproductively isolated

- But there are counterexamples: (Already known to Darwin)

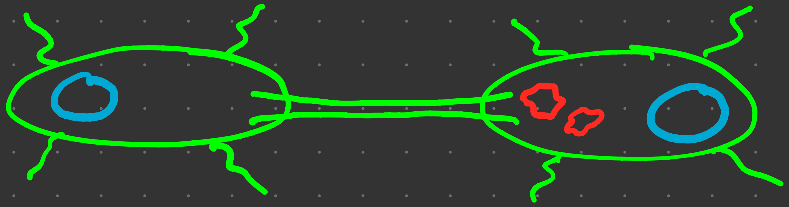


(E.g. some orchids)

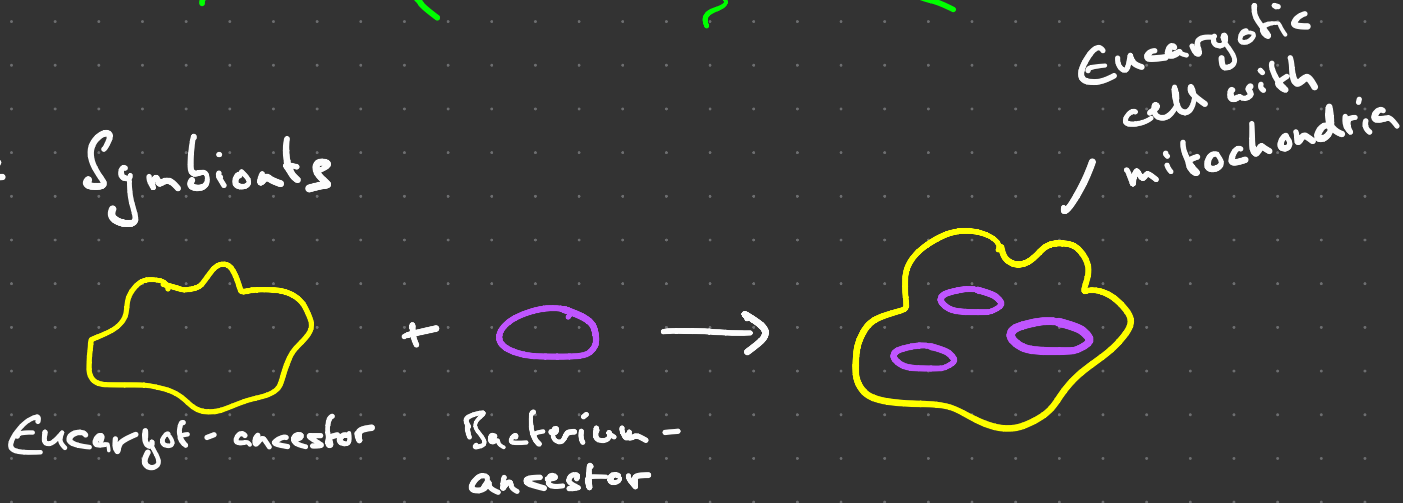
Fertile (!) hybrids of plants

Many more examples of "horizontal" evolution or "reticulate events"

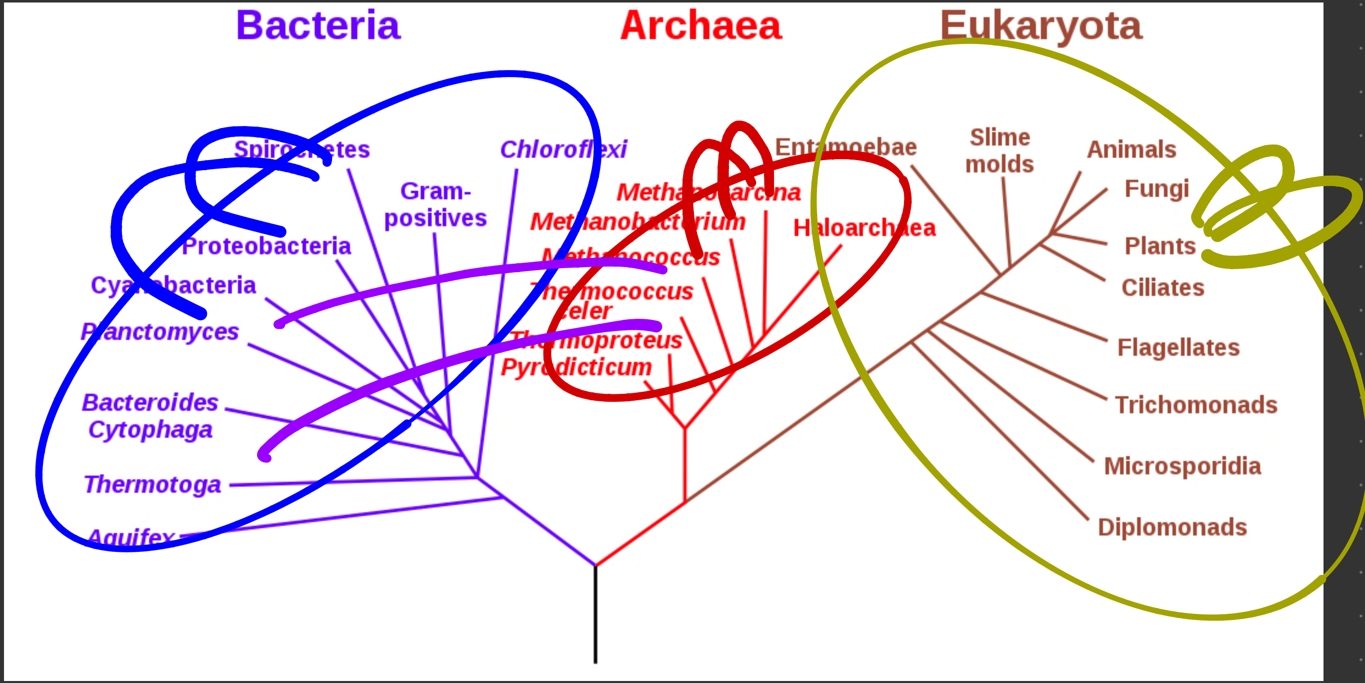
- "Conjugation" of bacteria & archaea



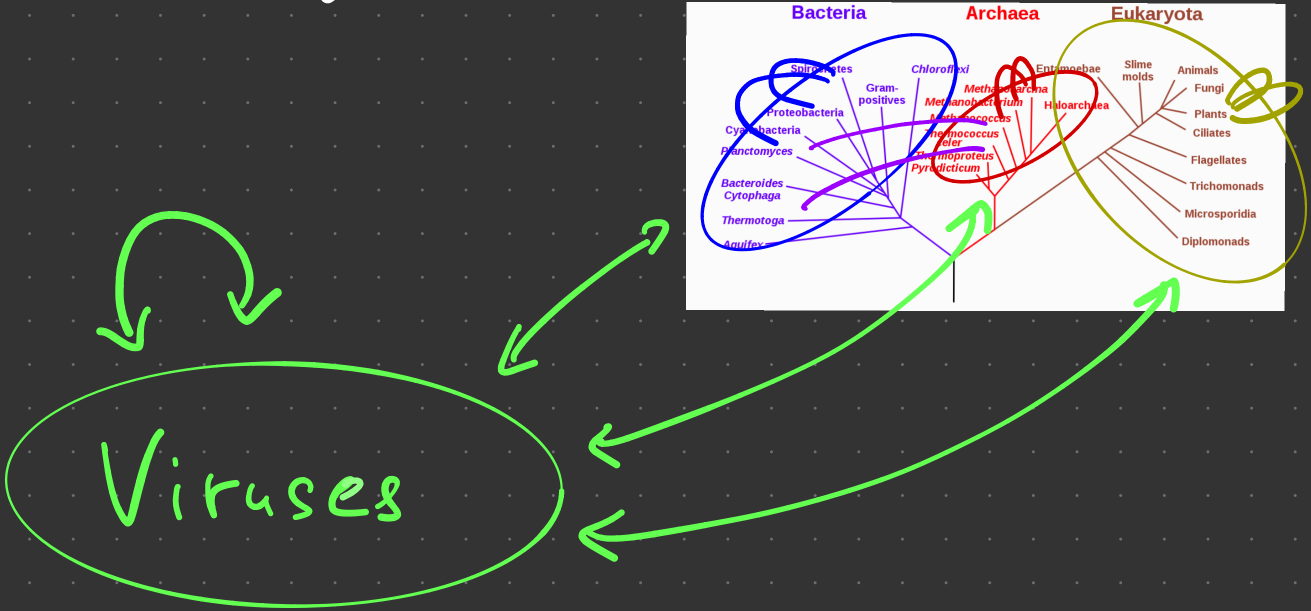
- Symbionts



Tree of Life ?

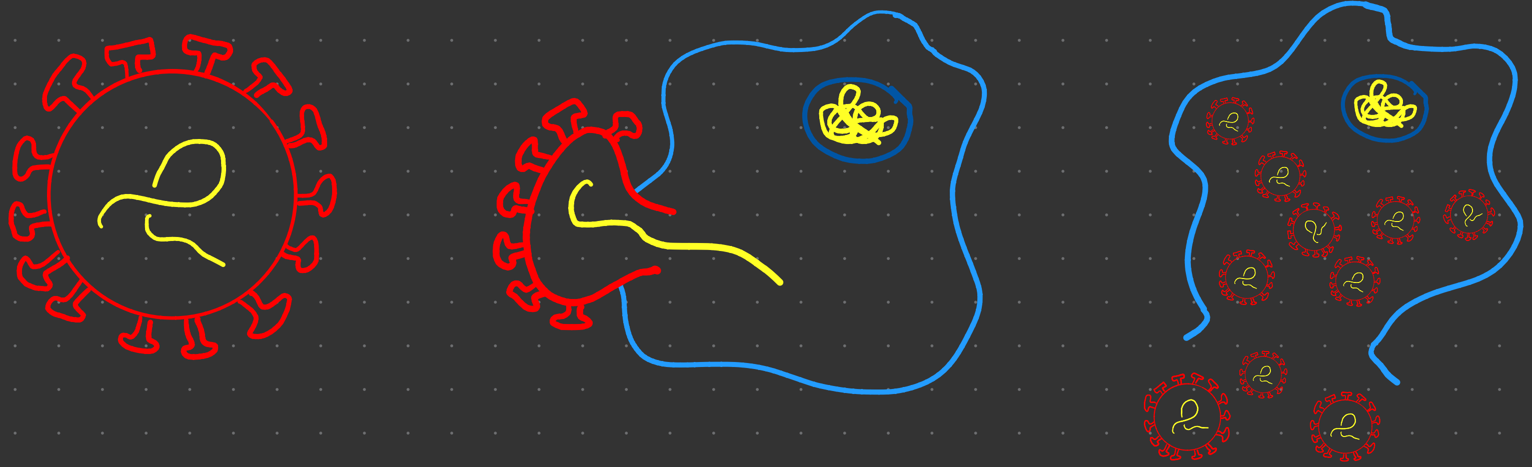

Bacteria
Archaea
Eukaryota
Spirochetes
Chloroflexi
Gram-positives
Proteobacteria
Cyanobacteria
Planctomyces
Bacteroides Cytophaga
Thermotoga
Aquifex
Methanosarcina
Methanobacterium
Haloarchaea
Methanococcus
Thermococcus celer
Thermoproteus
Pyrodicticum
Entamoebae
Slime molds
Animals
Fungi
Plants
Ciliates
Flagellates
Trichomonads
Microsporidia
Diplomonads

One more thing...



Reproduction of viruses

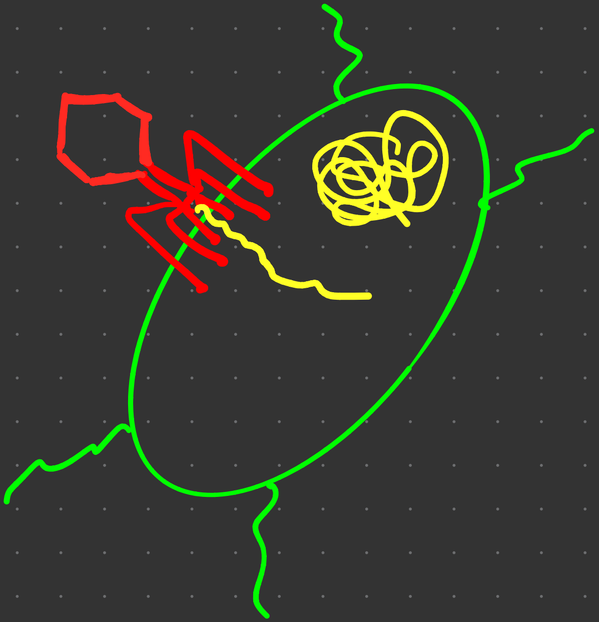
Viruses need a <u>host</u> to replicate

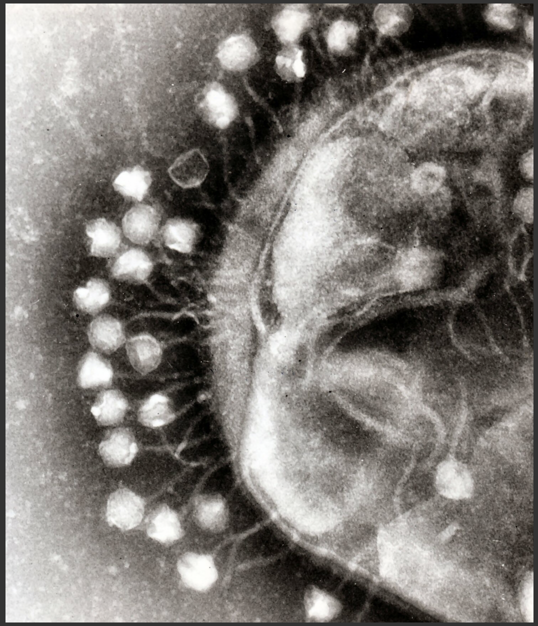


* If genome is built into gametes, it is inherited

⟹ 5-8% of human genome

Also for bacteria / archaea – "Bacteriophages"

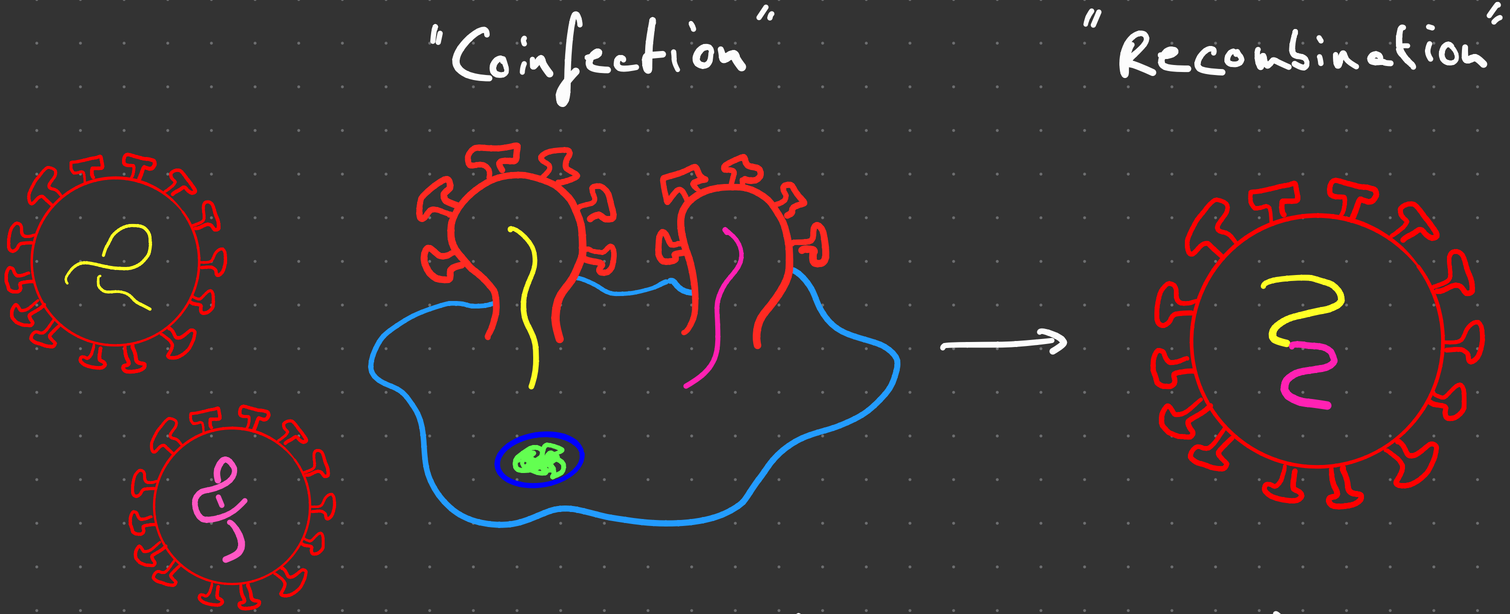




Source: Wiki

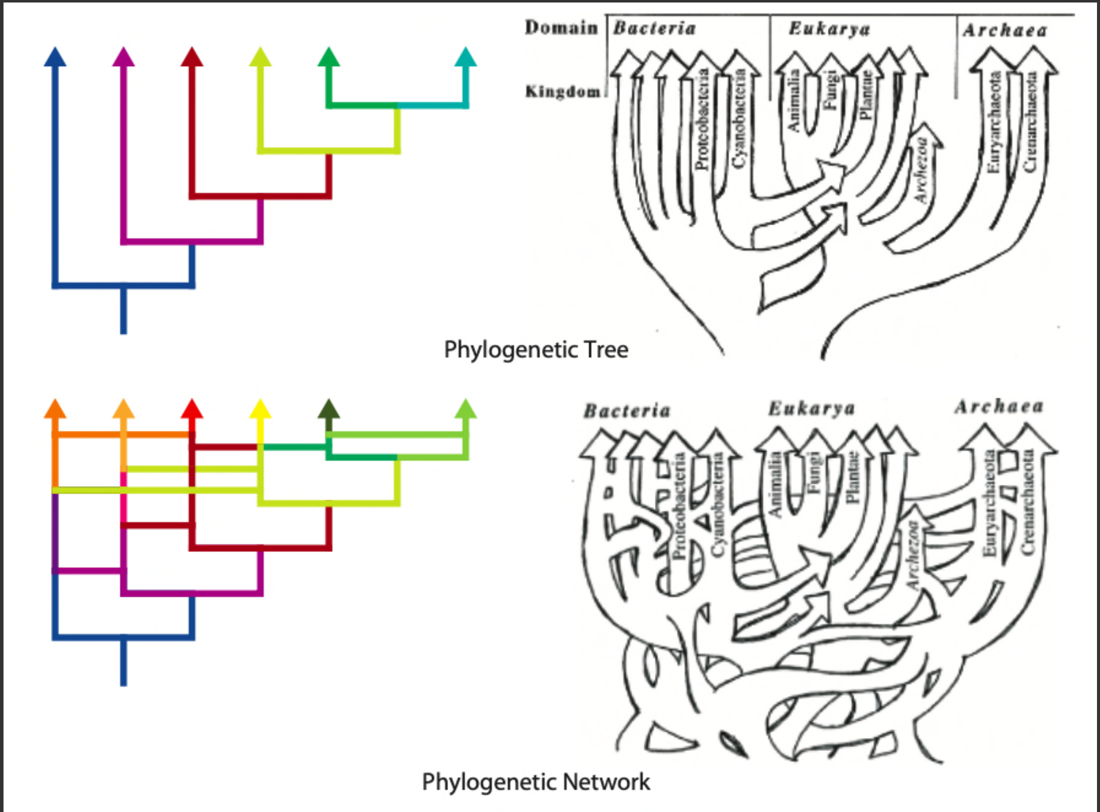
(Possibly responsible for the 2011 "EHEC" outbreak)

There is also mixing of genetic material between viruses



(see Example)

Instead of a tree, we need a network

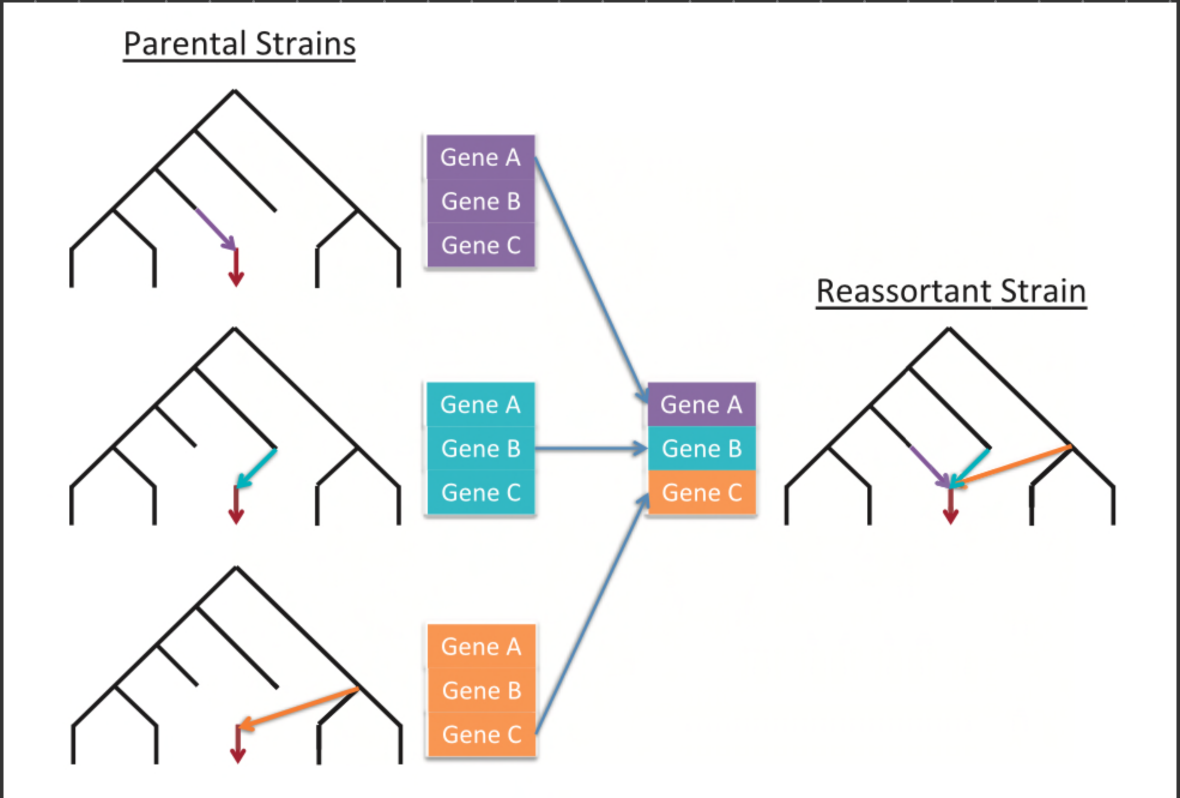


Phylogenetic Tree

Phylogenetic Network

Source: [2]

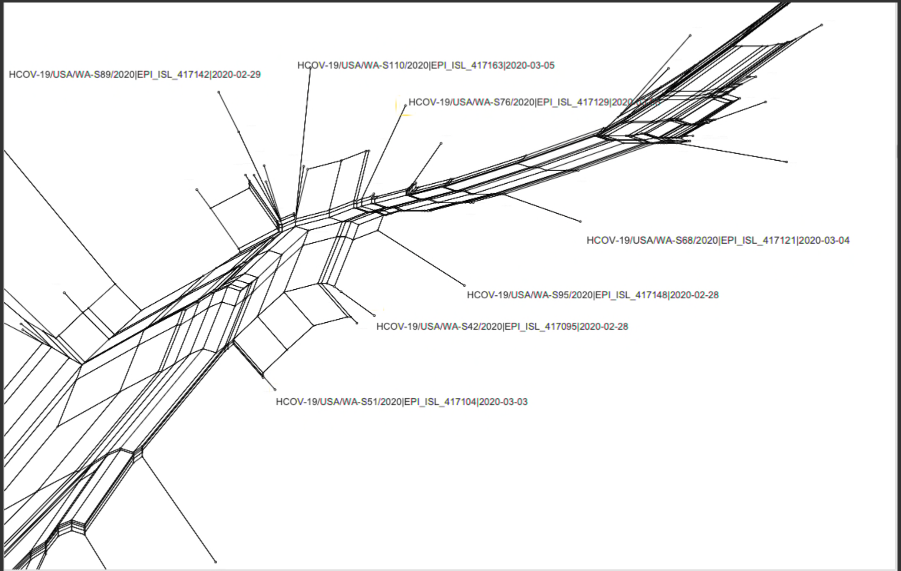
In principle, it is possible to construct the network



Source: [2]

However, this seems to be extremely difficult in real life, due to + computational complexity

+ biological interpretation issues


HCOV-19/USA/WA-S89/2020|EPI_ISL_417142|2020-02-29
HCOV-19/USA/WA-S110/2020|EPI_ISL_417163|2020-03-05
HCOV-19/USA/WA-S76/2020|EPI_ISL_417129|2020
HCOV-19/USA/WA-S68/2020|EPI_ISL_417121|2020-03-04
HCOV-19/USA/WA-S95/2020|EPI_ISL_417148|2020-02-28
HCOV-19/USA/WA-S42/2020|EPI_ISL_417095|2020-02-28
HCOV-19/USA/WA-S51/2020|EPI_ISL_417104|2020-03-03

← "Split-Network"

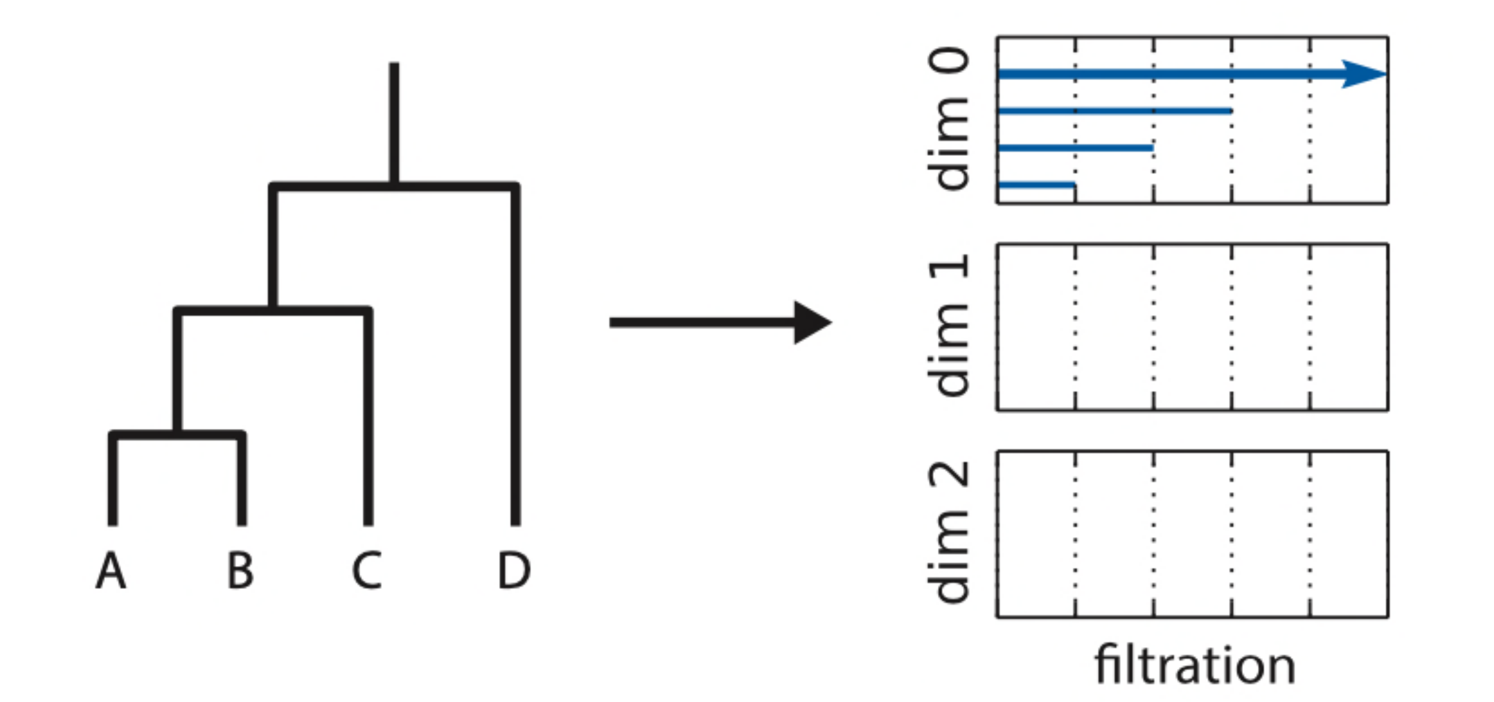
Hard to get relevant information

How Persistent Homology is able to help

Philosophy: * Evolution happens in a complicated, high-dimensional space (?) and we want to know its structure.

* How is this structure related to relevant information about evolution, e.g. type, scale & statistics of reticulate events.

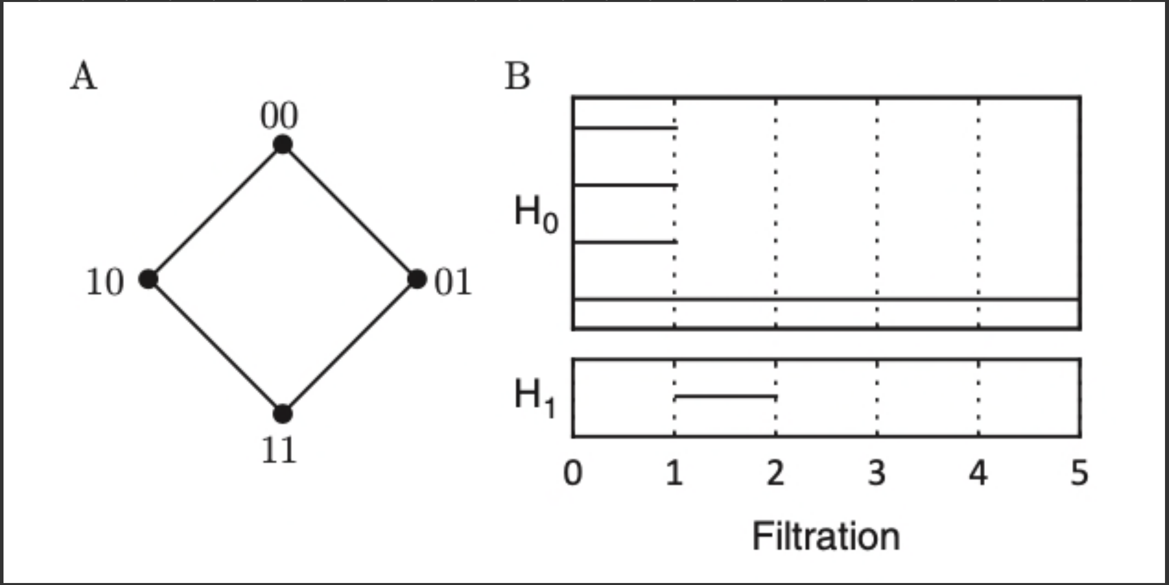
Hint: Reticulate events prevent contractibility of the network.



Source : [2]

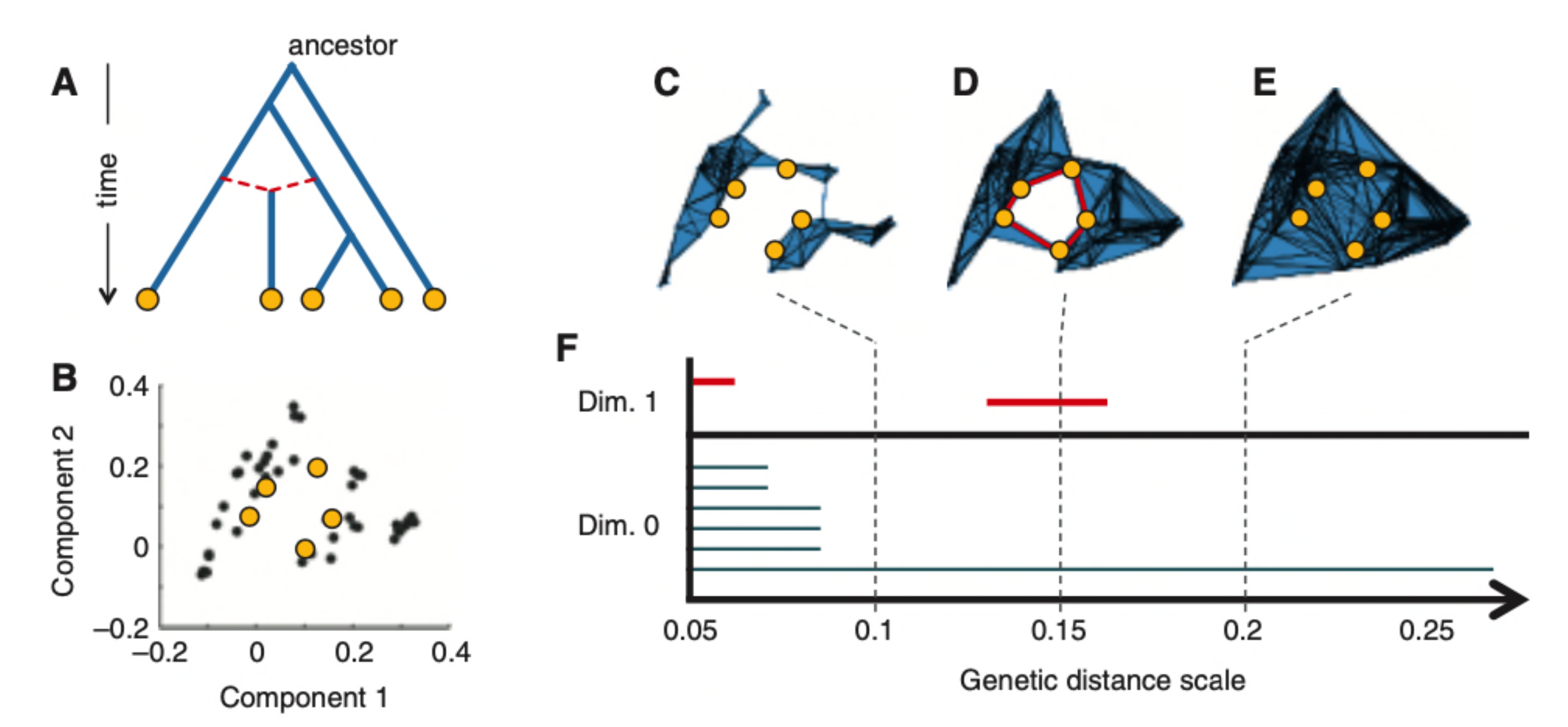
Simplest example

Suppose
- $\mathcal{A} = \{0,1\}$
- Every site only mutates once ("infinite-sites assumption")



Source: [2]

For a realistic example:



Source: [2]

Theorem (Chan, Carlsson, Rabadan)

(X, d_X) a tree-like finite metric space, $\varepsilon \geq 0$.

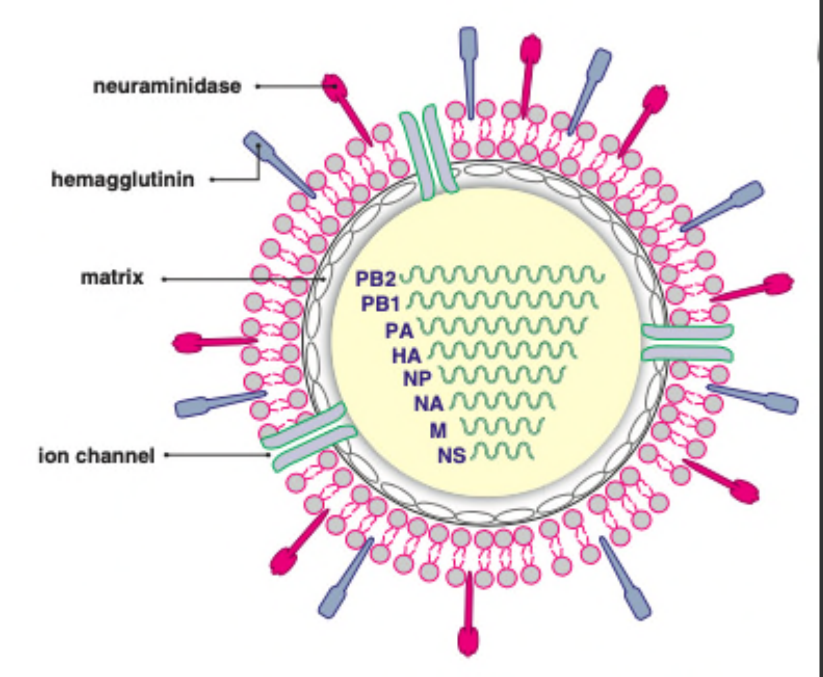
Then $H_i(VR_\varepsilon(X, d_X)) = \{0\}$ for $i \geqslant 0$.

$\Longrightarrow$ Topological obstruction: L^∞-norm on bars.

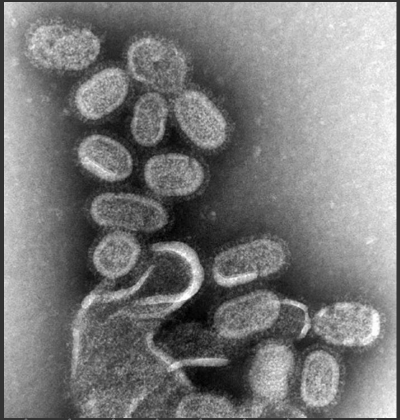
(Identification of "noise" by usual stability results – GH-metric bounds bottleneck-metric)

Example: Influenza

- * RNA virus
- * 8 "segments"
- * mutates very fast



Source: [2]

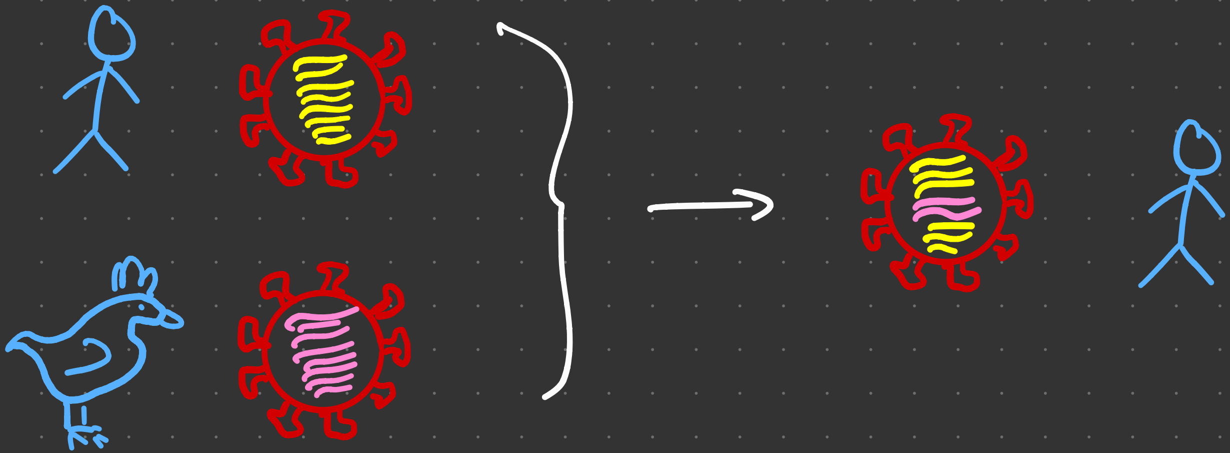


Source: Wiki

In principle two modes of evolution:

- vertically ~ 10^{-3} substitutions / nucleotide · year
- horizontally, through coinfections

Important feature: "Reassortment"

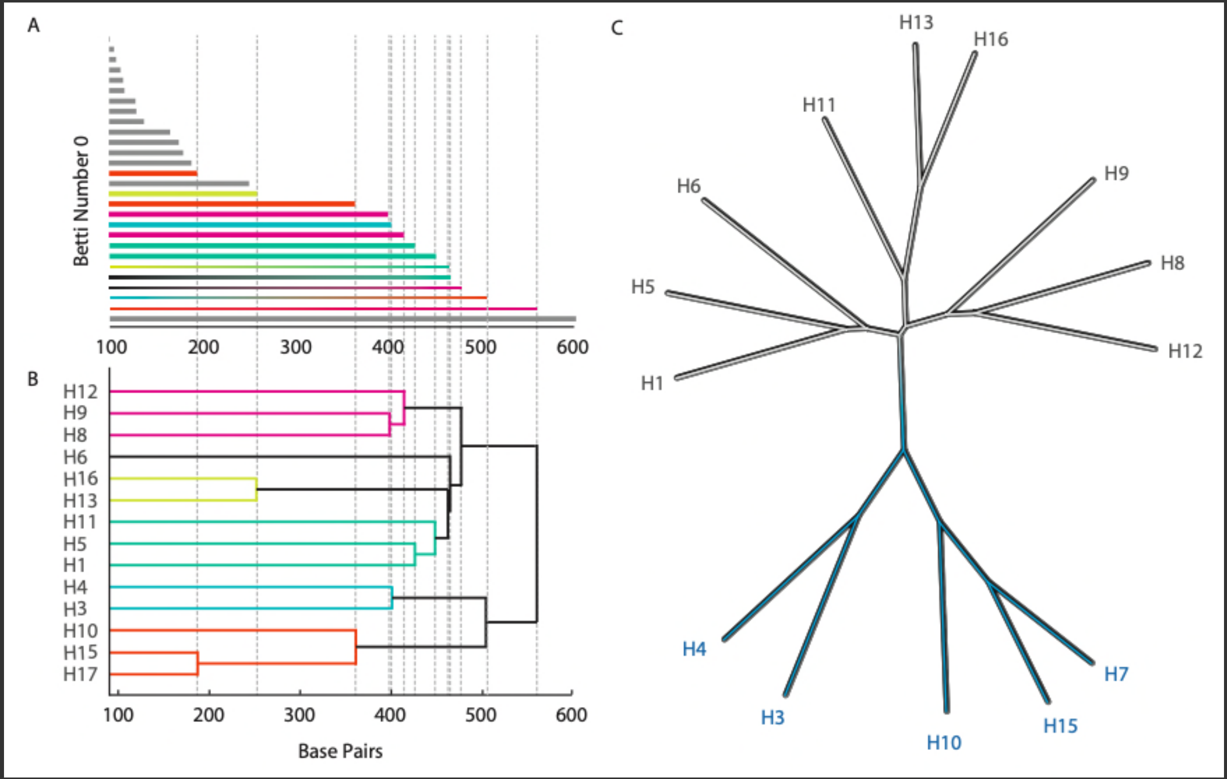


Questions concerning vaccines etc.

- Reassortment hotspots?
- Location?
- Rate?
- History?

⟹ Use TDA

For <u>one</u> segment:

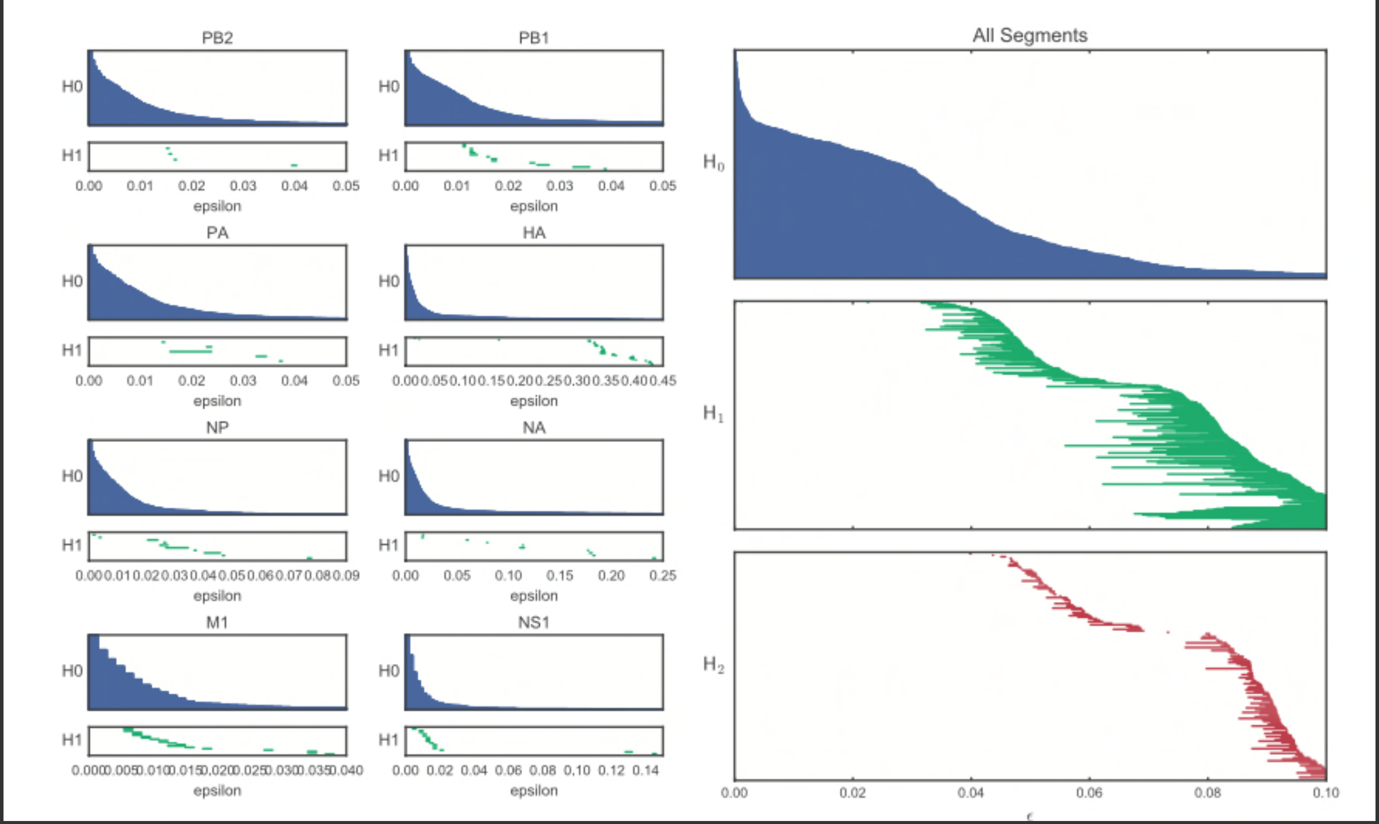


Source: [2]

Only relevant homology in dim. 0

→ no recombination

For the whole genome:



Source: [2]

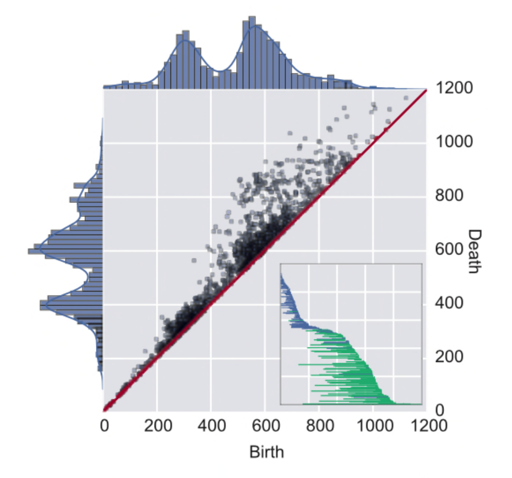
Barcodes provide useful information:

- * cycles correspond to reticulate events of various types
- * Statistics of cycles reveal reassortment hotspot preserving vital functions
- * Reassortment location and rates

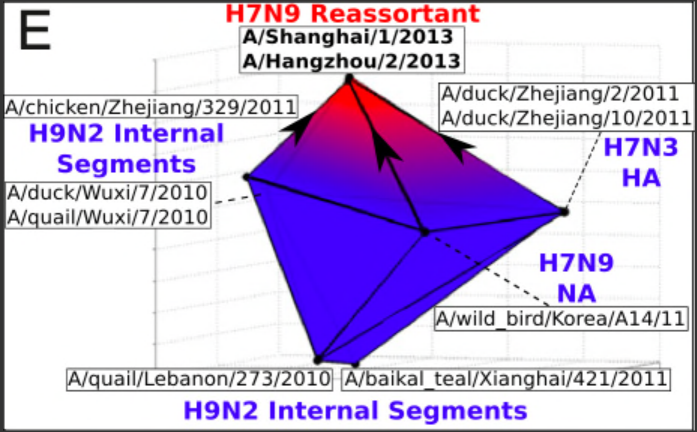
In particular, there is interesting information

... in the persistence diagram

... in higher dim. cycles



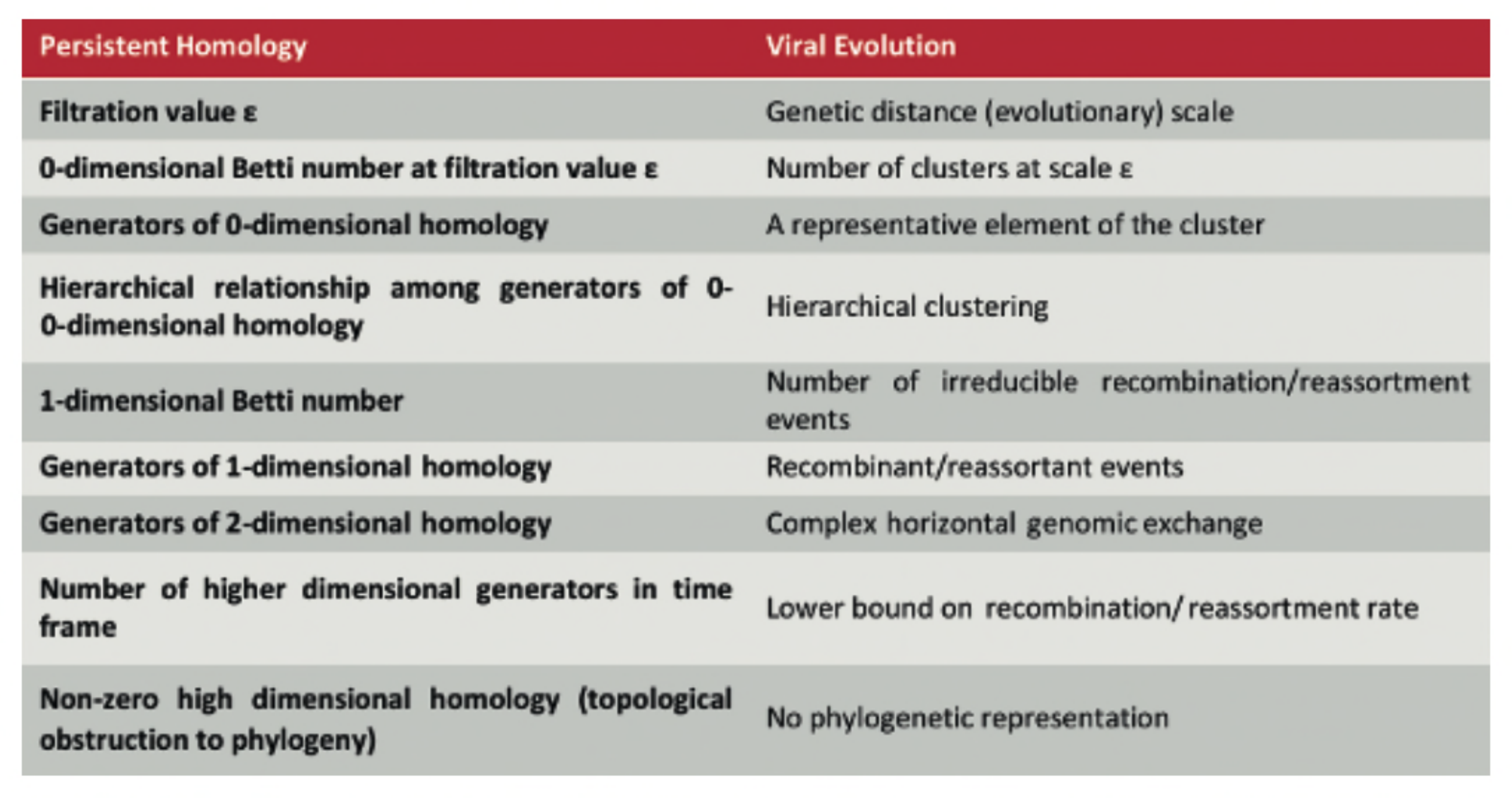
Source: [2]



Source: [2]

Intra- & Inter-subtype reassortments

Triple reassortment lead to 2013 H7N9 avian flu outbreak in China.



Persistent Homology	Viral Evolution
Filtration value ε	Genetic distance (evolutionary) scale
0-dimensional Betti number at filtration value ε	Number of clusters at scale ε
Generators of 0-dimensional homology	A representative element of the cluster
Hierarchical relationship among generators of 0-0-dimensional homology	Hierarchical clustering
1-dimensional Betti number	Number of irreducible recombination/reassortment events
Generators of 1-dimensional homology	Recombinant/reassortant events
Generators of 2-dimensional homology	Complex horizontal genomic exchange
Number of higher dimensional generators in time frame	Lower bound on recombination/ reassortment rate
Non-zero high dimensional homology (topological obstruction to phylogeny)	No phylogenetic representation

Source: [2]

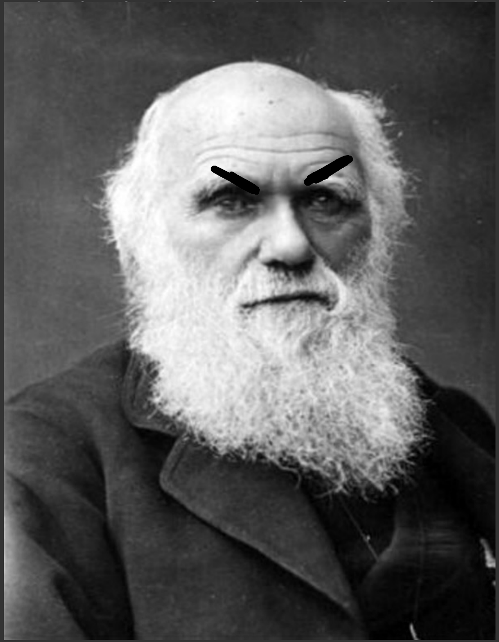
Summary

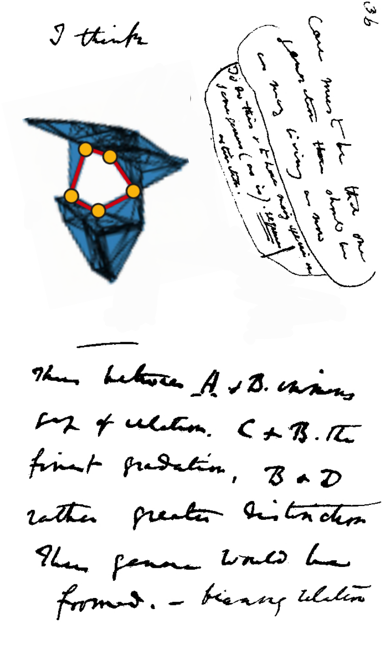
* Persistent homology provides a new way to think about evolution beyond trees & networks

* Barcodes and their statistics yield insights into evolutionary history on earth and pathogens

* Possible applications for vaccines, antibiotics and pandemics

Thank you!





36

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