Modeling sequence evolution

Olivier Gascuel

Directeur de Recherche au CNRS

Institut de Systématique, Evolution, Biodiversité (ISYEB)
Muséum National d'Histoire Naturelle

Académie des Sciences

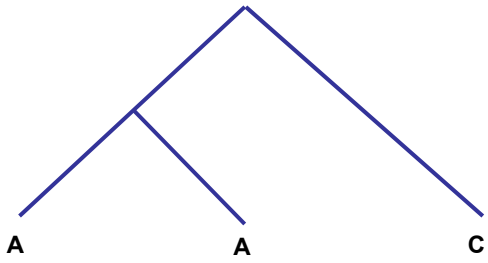
olivier.gascuel@mnhn.fr
https://isyeb.mnhn.fr/fr/annuaire/olivier-gascuel-7496

Olivier Gascuel – CNRS – Muséum National d'Histoire Naturelle – Académie des Sciences – Phylogenomics-Cesky-2024

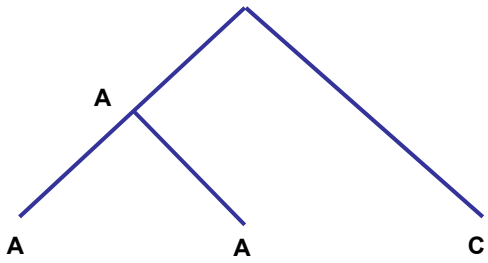
Modeling sequence evolution

- **Models for DNA**
  - Standard assumptions - Cartoon
  - The simplest RY Markov model – Mathematical basis
  - JC69, K2P, F81, HKY
  - GTR
- **Models for proteins**
  - JTT, WAG, LG and others
  - Options and estimation
- **Rates across sites models**
  - Gamma distributed rates
  - Free rates
- **Advanced models**
  - Mixtures and partitions
  - CAT, CXX (e.g. C60)
  - Heterotach, non-homogeneous, non-time reversible, others…
- **Model selection**

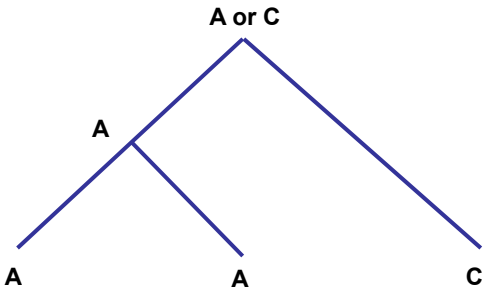
Modelling sequence (character) evolution



Modelling sequence (character) evolution

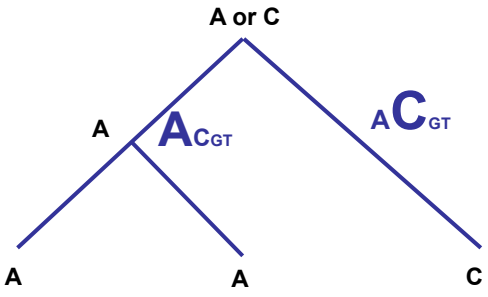


Modelling sequence (character) evolution



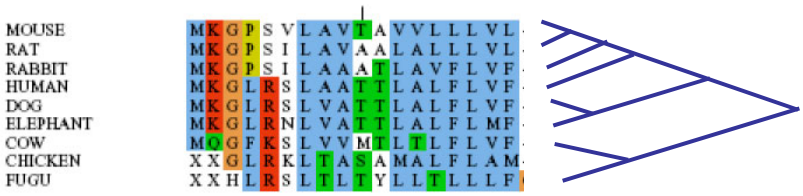
Parsimony

Modelling sequence (character) evolution

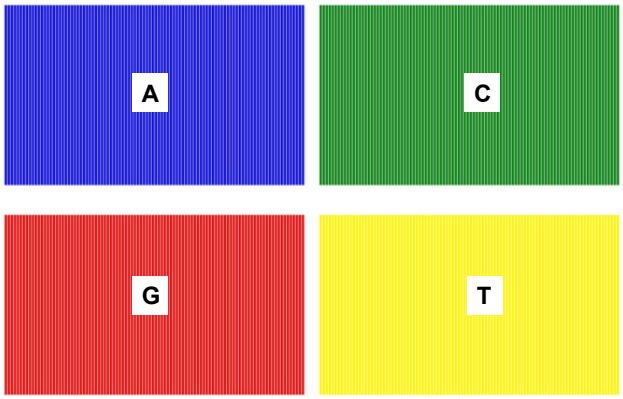


Parsimony **Probabilistic modelling**

Modeling sequence evolution: standard assumptions



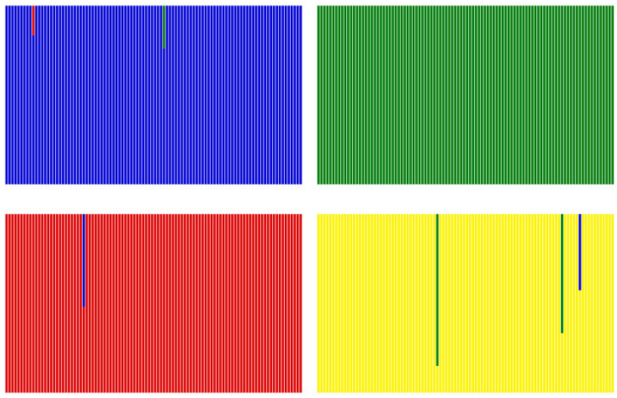
- Evolution is **independent among lineages**
- Evolution is **memory-less (Markov model)**
- The sites evolve **independently** and **identically**
- Models are **time reversible**
- Models are **time homogeneous** and **stationary**
- **Gaps** are treated as **unknowns**



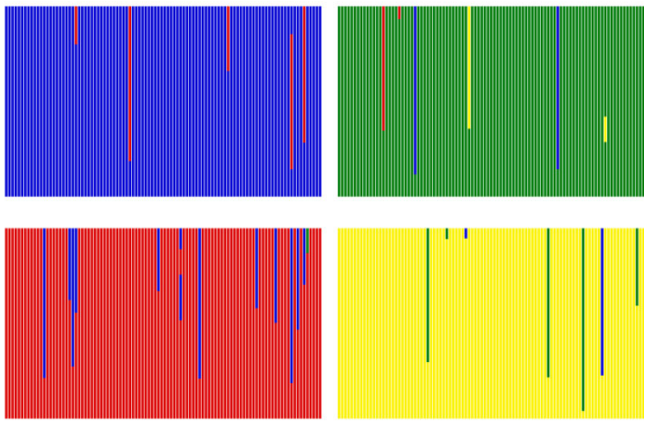
$\mu t = 0.0$ (mutations per site)

μ : mutation rate (number of mutations per site per time unit)

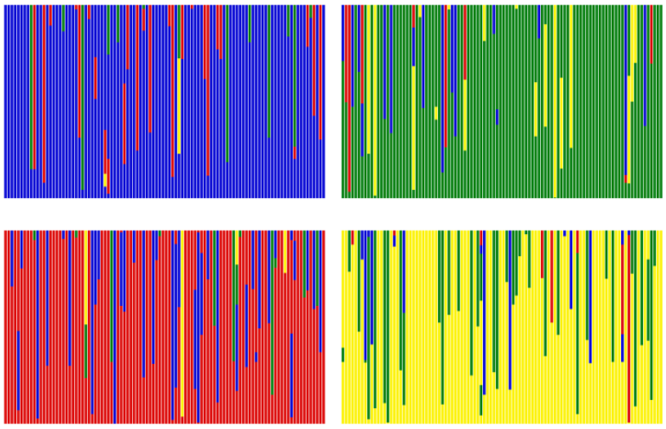
t : time



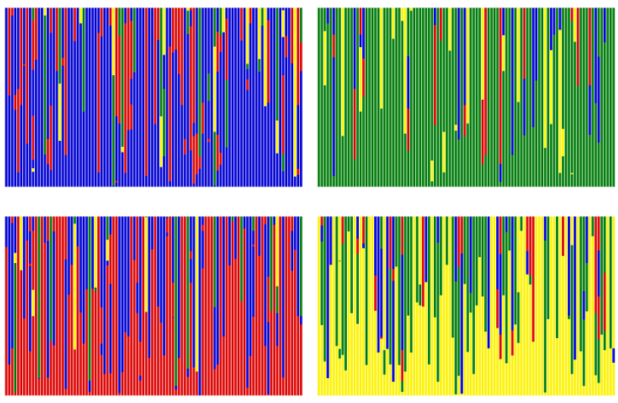
$\mu t = 0.01$ (mutations per site)



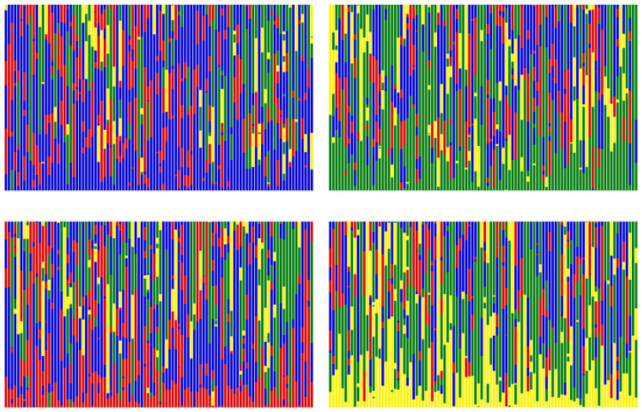
$\mu t = 0.10$ (mutations per site)



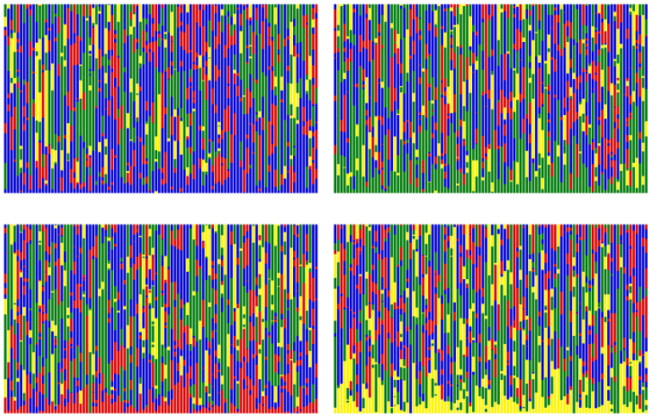
$\mu t = 0.50$ (mutations per site)



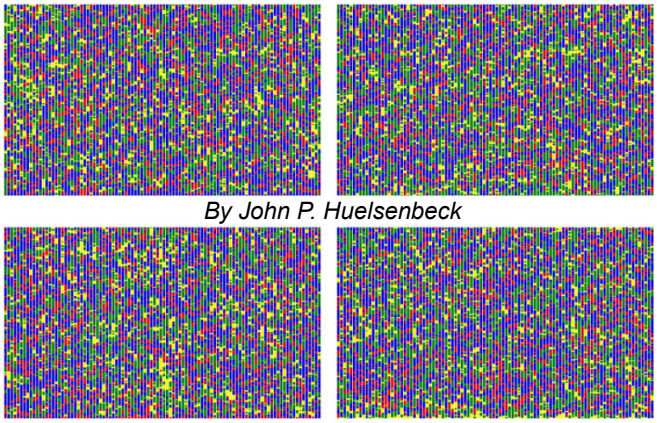
$\mu t = 1.00$ (mutations per site)



$\mu t = 5.00$ (mutations per site)



$\mu t = 10.00$ (mutations per site)



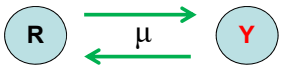
By John P. Huelsenbeck

$\mu t = 100.00$ (mutations per site)

The nucleotide frequencies are independent of the starting points!

They are equal to the "stationary distribution" of the Markov process.

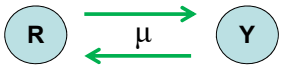
The simplest RY (0,1) symmetrical Markov, time continuous model



R Y R .....

expected number of mutations = μt

The simplest RY (0,1) symmetrical Markov, time continuous model



The rate matrix: $Q = \begin{pmatrix} - & \mu \\ \mu & - \end{pmatrix}$

The matrix of probability changes: $P(t) = \begin{pmatrix} \frac{1}{2}\left(1+e^{-2\mu t}\right) & \frac{1}{2}\left(1-e^{-2\mu t}\right) \\ \frac{1}{2}\left(1-e^{-2\mu t}\right) & \frac{1}{2}\left(1+e^{-2\mu t}\right) \end{pmatrix}$

$P(t) = e^{Qt}$

The equilibrium distribution: $\pi_R = \pi_Y = \frac{1}{2}$

The simplest RY (0,1) symmetrical Markov, time continuous model

The rate matrix: $Q = \begin{pmatrix} - & \mu \\ \mu & - \end{pmatrix}$

The matrix of probability changes: $P(t) = \begin{pmatrix} \frac{1}{2}\left(1+e^{-2\mu t}\right) & \frac{1}{2}\left(1-e^{-2\mu t}\right) \\ \frac{1}{2}\left(1-e^{-2\mu t}\right) & \frac{1}{2}\left(1+e^{-2\mu t}\right) \end{pmatrix}$

The equilibrium distribution: $\pi_R = \pi_Y = \frac{1}{2}$

This model is time-reversible: $\pi_X P_{X\to Y}(t) = \pi_Y P_{Y\to X}(t)$

We assume stationnarity (frequencies of R and Y are nearly equal)

The simplest RY (0,1) symmetrical Markov, time continuous model

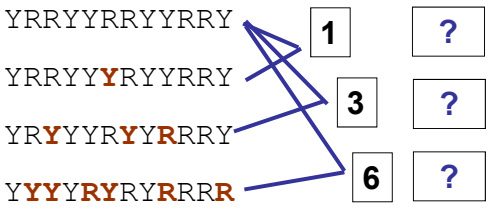
The rate matrix: $Q = \begin{pmatrix} - & \mu \\ \mu & - \end{pmatrix}$

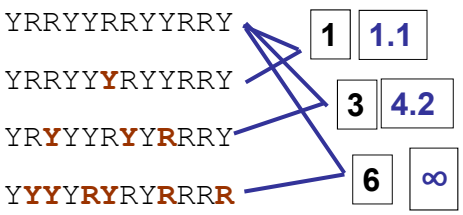
The matrix of probability changes: $P(t) = \begin{pmatrix} \frac{1}{2}\left(1+e^{-2\mu t}\right) & \frac{1}{2}\left(1-e^{-2\mu t}\right) \\ \frac{1}{2}\left(1-e^{-2\mu t}\right) & \frac{1}{2}\left(1+e^{-2\mu t}\right) \end{pmatrix}$

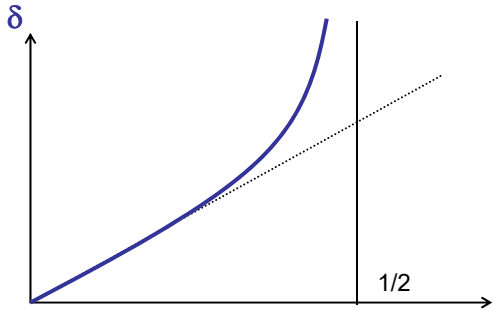
The equilibrium distribution: $\pi_R = \pi_Y = \frac{1}{2}$

Within this model the evolutionary distance (expected number of substitutions per site) between two sequences is:

$$\delta = -\frac{1}{2}\log\left(1-2P_{S_i \neq S'_i}\right)$$







$$\delta = -\frac{1}{2}\log\left(1 - 2P_{S_i \neq S'_i}\right)$$

Jukes and Cantor model (JC69) for DNA

M =

↗	A	T	C	G
A	–	μ	μ	μ
T	μ	–	μ	μ
C	μ	μ	–	μ
G	μ	μ	μ	–

Eq. (1/4, 1/4, 1/4, 1/4)

The Jukes & Cantor model has been historically the first one to be introduced.
Justification: simplicity

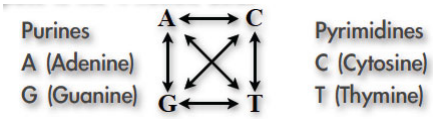
Kimura 2-parameter (K2P) model for DNA

$M =$

↗	A	T	C	G
A	–	β	β	α
T	β	–	α	β
C	β	α	–	β
G	α	β	β	–

Eq. (1/4, 1/4, 1/4, 1/4)

Kimura's 2-parameter model aims at reflecting the fact that transitions are (~twice) more frequent than transitions



Felsenstein 1981 (F81) model for DNA

$M =$

↗	A	T	C	G
A	–	$\mu \pi_T$	$\mu \pi_C$	$\mu \pi_G$
T	$\mu \pi_A$	–	$\mu \pi_C$	$\mu \pi_G$
C	$\mu \pi_A$	$\mu \pi_T$	–	$\mu \pi_G$
G	$\mu \pi_A$	$\mu \pi_T$	$\mu \pi_C$	–

Eq. $(\pi_A, \pi_T, \pi_C, \pi_G)$

Felsenstein's 1981 model allows for any arbitrary set of equilibrium frequencies.

Hasegawa, Kishino,Yano (HKY) model for DNA

$M =$

↗	A	T	C	G
A	–	$\beta\,\pi_T$	$\beta\,\pi_C$	$\alpha\,\pi_G$
T	$\beta\,\pi_A$	–	$\alpha\,\pi_C$	$\beta\,\pi_G$
C	$\beta\,\pi_A$	$\alpha\,\pi_T$	–	$\beta\,\pi_G$
G	$\alpha\,\pi_A$	$\beta\,\pi_T$	$\beta\,\pi_C$	–

Eq. $(\pi_A, \pi_T, \pi_C, \pi_G)$

The HKY model is a way to incorporate both transition/transversion bias and an arbitrary set of equilibrium frequencies. F84 is very similar. **Both capture the two main aspects of DNA evolution.**

25

Most general time-reversible (GTR) model for DNA

$M =$

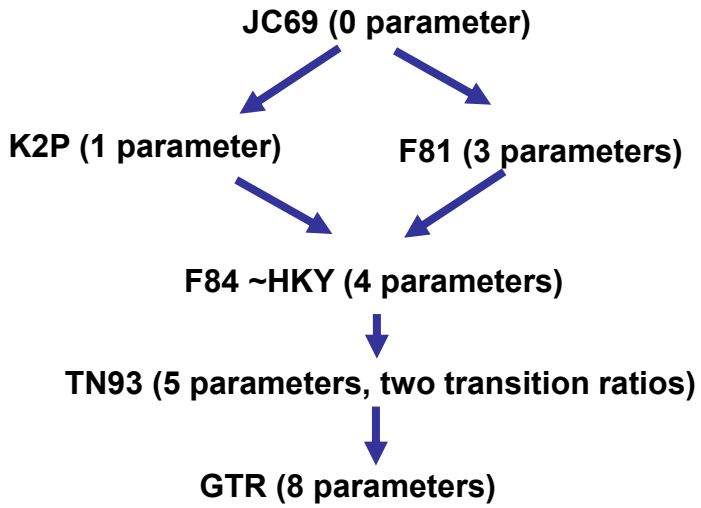
↗	A	T	C	G
A	–	$\rho_{AT}\pi_T$	$\rho_{AC}\pi_C$	$\rho_{AG}\pi_G$
T	$\rho_{AT}\pi_A$	–	$\rho_{CT}\pi_C$	$\rho_{GT}\pi_G$
C	$\rho_{AC}\pi_A$	$\rho_{CT}\pi_T$	–	$\rho_{CG}\pi_G$
G	$\rho_{AG}\pi_A$	$\rho_{GT}\pi_T$	$\rho_{CG}\pi_C$	–

Eq. $(\pi_A, \pi_T, \pi_C, \pi_G)$, the ρ parameters are symmetrical and named exchangeabilities

Model to be preferred when having enough data

26

DNA model (nesting) hierarchy



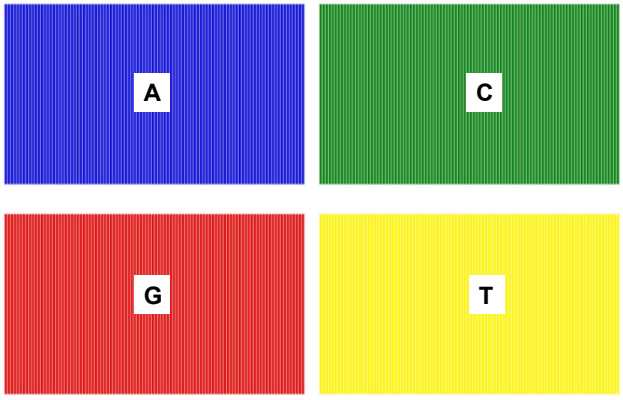
All parameters are estimated from the alignment being analysed

27

$$\mathbf{Q} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

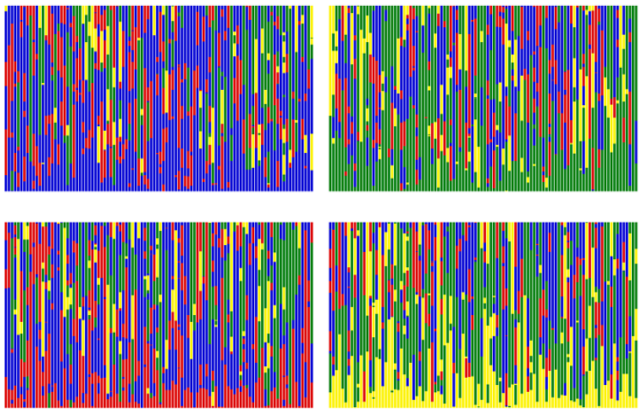
$$P(t) = e^{Qt} = I + Qt + \frac{1}{2}(Qt)^2 + \frac{1}{3!}(Qt)^3 \ldots . \frac{1}{n!}(Qt)^n \ldots$$

28



$$\mathbf{P}(0.0) = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}$$

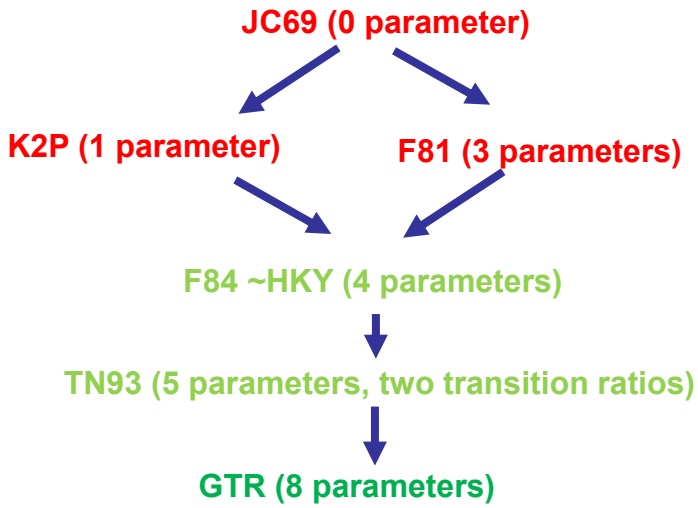
29



$$\mathbf{P}(5.0) = \begin{pmatrix} 0.138 & 0.188 & 0.494 & 0.180 \\ 0.138 & 0.190 & 0.492 & 0.181 \\ 0.137 & 0.187 & 0.497 & 0.178 \\ 0.138 & 0.188 & 0.494 & 0.180 \end{pmatrix}$$

30

DNA model (nesting) hierarchy



All models are wrong, some are (super) useful

31

Models for proteins (JTT, WAG, LG… generalist)

We use GTR models, which involves 208 free parameters (190-1 symmetrical exchangeabilities, 20-1 amino-acid frequencies)

Except with very large alignments, it's not possible to estimate so many parameters from the data analyzed

These models are estimated from large sets of alignments (LG: several thousands of alignments, comprizing millions of residues)

32

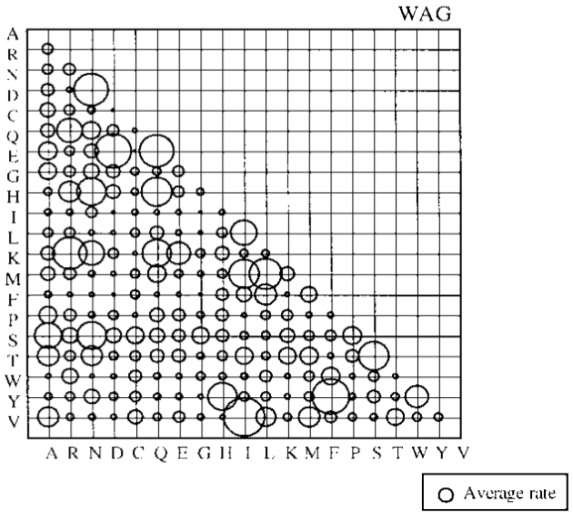
Models for proteins (JTT, WAG, LG… generalist)

Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
-																			
0.55	-																		
0.51	0.64	-																	
0.74	0.15	5.43	-																
1.03	0.53	0.27	0.03	-															
0.91	3.04	1.54	0.62	0.10	-														
1.58	0.44	0.95	6.17	0.02	5.47	-													
1.42	0.58	1.13	0.87	0.31	0.33	0.57	-												
0.32	2.14	3.96	0.93	0.25	4.29	0.57	0.25	-											
0.19	0.19	0.55	0.04	0.17	0.11	0.13	0.03	0.14	-										
0.40	0.50	0.13	0.08	0.38	0.87	0.15	0.06	0.50	3.17	-									
0.91	5.35	3.01	0.48	0.07	3.89	2.58	0.37	0.89	0.32	0.26	-								
0.89	0.68	0.20	0.10	0.39	1.55	0.32	0.17	0.40	4.26	4.85	0.93	-							
0.21	0.10	0.10	0.05	0.40	0.10	0.08	0.05	0.68	1.06	2.12	0.09	1.19	-						
1.44	0.68	0.20	0.42	0.11	0.93	0.68	0.24	0.70	0.10	0.42	0.56	0.17	0.16	-					
3.37	1.22	3.97	1.07	1.41	1.03	0.70	1.34	0.74	0.32	0.34	0.97	0.49	0.55	1.61	-				
2.12	0.55	2.03	0.37	0.51	0.86	0.82	0.23	0.47	1.46	0.33	1.39	1.52	0.17	0.80	4.38	-			
0.11	1.16	0.07	0.13	0.72	0.22	0.16	0.34	0.26	0.21	0.67	0.14	0.52	1.53	0.14	0.52	0.11	-		
0.24	0.38	1.09	0.33	0.54	0.23	0.20	0.10	3.87	0.42	0.40	0.13	0.43	6.45	0.22	0.79	0.29	2.49	-	
2.01	0.25	0.20	0.15	1.00	0.30	0.59	0.19	0.12	7.82	1.80	0.31	2.06	0.65	0.31	0.23	1.39	0.37	0.31	-
8.66	4.40	3.91	5.70	1.93	3.67	5.81	8.33	2.44	4.85	8.62	6.20	1.95	3.84	4.58	6.95	6.10	1.44	3.53	7.09

LG model for proteins

33

Models for proteins (JTT, WAG, LG… generalist)



The area of each bubble represents the symmetrical exchangeability (ρ_{ij}) for the replacement of amino acid *i* by amino acid *j* or vice versa.

34

Models for proteins: Main options

-F : use the default, average model frequencies

+F : use the amino-frequencies estimated from your alignment (+ 19 parameters, for large data sets)

+FO : use the amino-frequencies estimated by ML from your alignment and tree (+ 19 param, large data sets)

Models for proteins (specific)

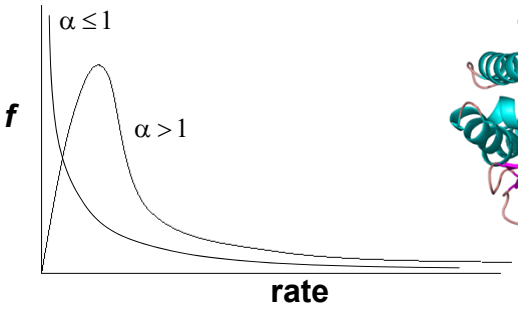
Specific matrices have been estimated for mitochondria (e.g. MtMam), protein groups (e.g. membrane), and species (e.g. MtArt, HIV, FLU…).

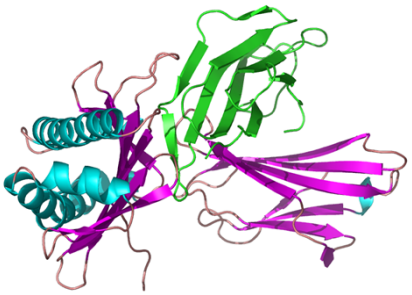
Use ProtTest or SMS to select the most appropriate for your dataset (or ModelFinder… etc.). Super useful!

Several methods, programs and servers are available for estimating new models, dedicated to your favorite protein groups (http://www.atgc-montpellier.fr/ReplacementMatrix/, CherryML…), when having many big MSAs…

Rates accross sites models

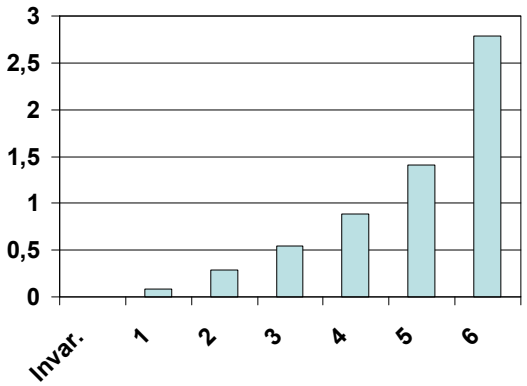
- It is usually assumed that sites evolve at different rates, which is modelled using a (discrete) gamma distribution, or a gamma plus invariant distribution, or a "free rate" model.





Rates accross sites models

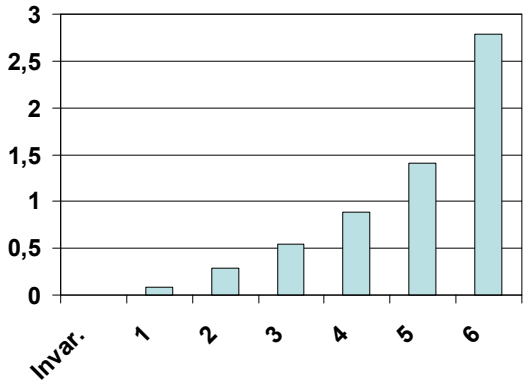
- It is usually assumed that sites evolve at different rates, which is modelled using a (discrete) gamma distribution (1 parameter), or a gamma plus invariant distribution (2 parameters), or a "free rate" model (12 parameters with 7 site categories).



Computing time 6-10 times slower!

Rates accross sites models Super useful!

- It is usually assumed that sites evolve at different rates, which is modelled using a (discrete) gamma distribution (1 parameter), or a gamma plus invariant distribution (2 parameters), or a "free rate" model (12 parameters with 7 site categories).



$\alpha = 1$
6 rate categories

Computing time
6-10 times slower!

39

Mixture and site-partition models

- Actually, we know that sites are evolving under different constraints (e.g. codon positions in DNA, buried/exposed residues in proteins).

- Partition models: we have some a priori knowledge on site categorization (e.g. codon position, 3D structure of the protein). Then, we use different models for each category.

- We do not know site categories (e.g. fast/slow sites, 3D structure). Then, we use a mixture where each site is analysed using all models within a predefined set of models (4 rate categories, LG4X, buried/exposed, EHO…). The likelihood of each site is the weighted sum of the likelihood for each model. We are able to predict the site category a posteriori.

40

CAT and CXX (e.g. C60) mixture models

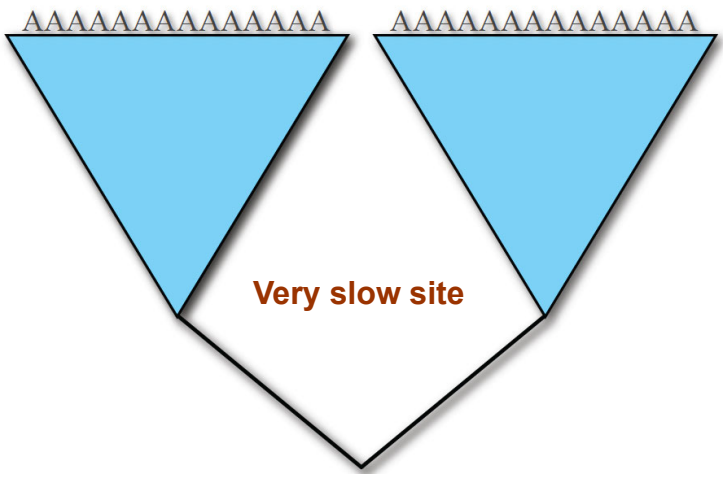
Lartillot Philippe MBE 2004 PhyloBayes, Le OG Lartillot Bioinformatics 2008 PhyML

- With highly divergent proteins, we observe that AA present on a given (fast) site are limited to a few possibilities (e.g. I/V), while Markov models predict that nearly all AAs should be observed.
- These saturated sites perturb tree inference and tend to produce long branch attraction artefacts.
- We use a mixture containing many (e.g. 60) simple F81-like models, that are defined by profiles (i.e. AA frequencies).
- These profiles are combined with gamma distributed rates.
- Model: *n* profiles (estimated from data in a Bayesian setting, once for all from huge databases in ML), *n* profile probabilities, gamma param.
- **Good results with deep phylogenies;** can be combined with JTT or LG; **computationally very heavy**…
- "**No overfitting**" *Banos et al Systematic Biology 2023*

41

Covarion, heterotach, Markov-modulated models

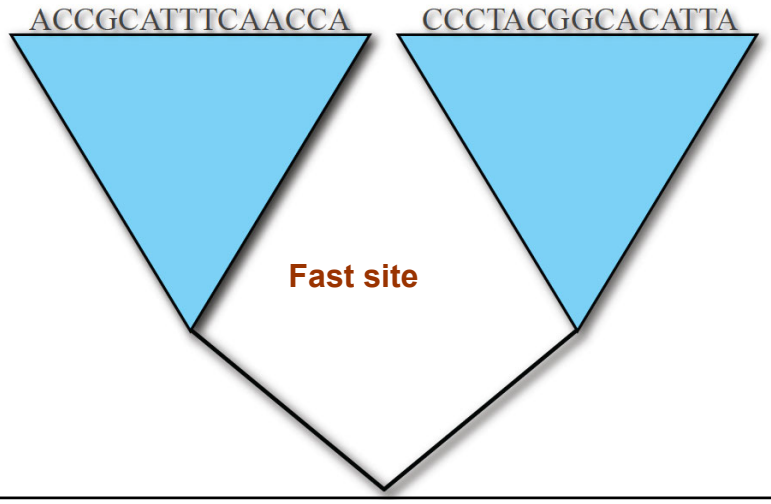
Heterotachy entails variation in the process of nucleotide substitution at individual sites in the sequence over the tree



42

Covarion, heterotach, Markov-modulated models

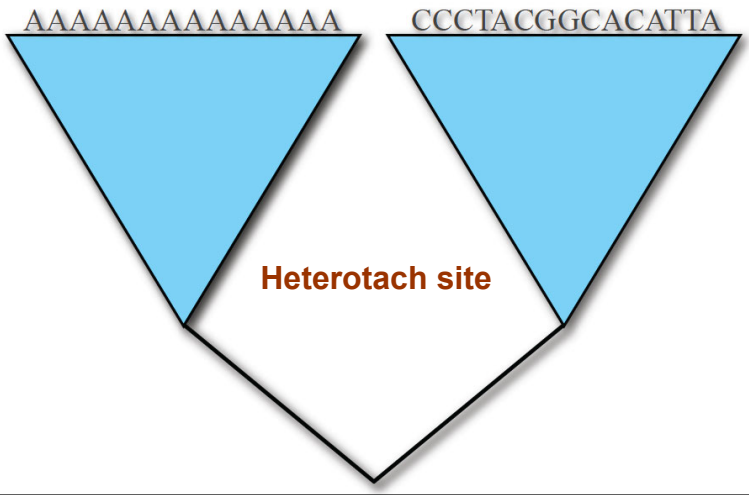
Heterotachy entails variation in the process of nucleotide substitution at individual sites in the sequence over the tree



43

Covarion, heterotach, Markov-modulated models

Heterotachy entails variation in the process of nucleotide substitution at individual sites in the sequence over the tree



44

Covarion, heterotach, Markov-modulated models

Heterotachy entails variation in the process of nucleotide substitution at individual sites in the sequence over the tree

$$Q = \begin{pmatrix} - & 0 & 0 & 0 & \lambda_{01} & 0 & 0 & 0 \\ 0 & - & 0 & 0 & 0 & \lambda_{01} & 0 & 0 \\ 0 & 0 & - & 0 & 0 & 0 & \lambda_{01} & 0 \\ 0 & 0 & 0 & - & 0 & 0 & 0 & \lambda_{01} \\ \lambda_{10} & 0 & 0 & 0 & - & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ 0 & \lambda_{10} & 0 & 0 & r_{AC}\pi_A & - & r_{CG}\pi_G & r_{CT}\pi_T \\ 0 & 0 & \lambda_{10} & 0 & r_{AG}\pi_A & r_{CG}\pi_C & - & \pi_T \\ 0 & 0 & 0 & \lambda_{10} & r_{AT}\pi_A & r_{CT}\pi_C & \pi_G & - \end{pmatrix}$$

$$Q = \left(\begin{array}{c|c} \text{Process is off (no substitutions are possible)} & \text{Switching from off to on} \\ \hline \text{Switching from on to off} & \text{Process is on (substitutions may occur)} \end{array}\right)$$

45

And more…….

- Codon models: dN/dS, empirical…
- Doublet models for RNA
- Non-homogeneous, non-time reversible…
- Many others……
- Some are useful!

46

Maximum-likelihood principle

Given *M* the model with parameters θ and *D* the data, we maximize:

$$L(M,\theta \mid D) = \Pr(D \mid M,\theta)$$

that is, the probability of generating *D* given *M* and θ.

A gold standard in statistics

- Asymptotically unbiased, minimum variance estimators
- Likelihood functions can be used to test hypotheses about models and parameters, typically using likelihood ratio tests (LRT)

47

LRT: a simple example with coin flipping

We flip a coin 10 times and observe 8 heads and 2 tails

Let p be the probability of head (and 1-p that of tail)

Two nested models:

M0 fair coin p = 0.5 — LK0 =

M1 unfair coin — LK1 =

p is estimated by ML — p =

LK1 =

Log-likelihood ratio test -2 x Log [LK0/LK1] =

Chi2 with 1 degree of freedom — p-value =

can we reject M0 ?

48

LRT: a simple example with coin flipping

We flip a coin 10 times and observe 8 heads and 2 tails

Let p be the probability of head (and 1-p that of tail)

Two **nested** models:

M0 fair coin p = 0.5	LK0 = 0.5^{10} ~ 0.001 x 45
M1 unfair coin	LK1 =
p is estimated by ML	p =
	LK1 =

Log-likelihood ratio test -2 x Log [LK0/LK1] =

Chi2 with 1 degree of freedom p-value =

LRT: a simple example with coin flipping

We flip a coin 10 times and observe 8 heads and 2 tails

Let p be the probability of head (and 1-p that of tail)

Two **nested** models:

M0 fair coin p = 0.5	LK0 = 0.5^{10} ~ 0.001 x 45
M1 unfair coin	LK1 = $p^8 (1-p)^2$ x 45
p is estimated by ML	p = 8/10
	LK1 ~ 0.007 x 45

Log-likelihood ratio test -2 x Log [LK0/LK1] ~ 2xLog [7] ~ 1.7

Chi2 with 1 degree of freedom p-value ~0.2

we cannot reject M0

Do it again with 9 heads and 1 tail !!

Model selection

- The higher the likelihood of the data, the better the model
- But, with high number of parameters, we expect high likelihood values, because we have many degrees of freedom to fit the model to the data at hand.
- We then have to penalize the likelihood value to account for the number of parameters.
- With nested models (e.g. JC69-K2P-HKY-GTR) we can use the likelihood ratio test (LRT). But it's a parametric approach....

51

Olivier Gascuel – CNRS – Muséum National d'Histoire Naturelle – Académie des Sciences – Phylogenomics-Cesky-2024

Model selection

- The higher the likelihood of the data, the better the model
- But, with high number of parameters, we expect high likelihood values, because we have many degrees of freedom to fit the model to the data at hand.
- We then have to penalize the likelihood value to account for the number of parameters.
- In most cases we use AIC or BIC, to be minimized (thanks to software programs: ModelTest, ProtTest, SMS, ModelFinder…).

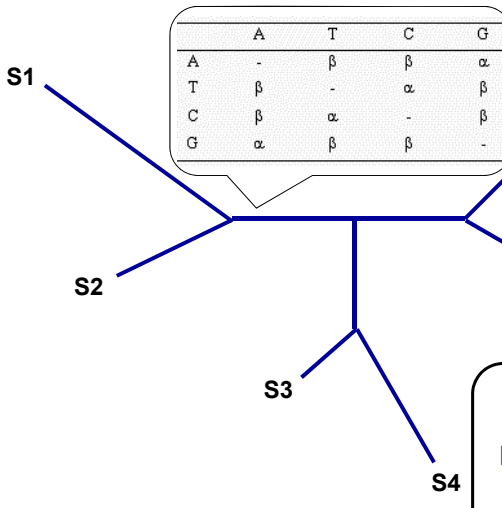
  AIC= - 2 LogLk(D/M) + 2 x #parameters

  BIC= - 2 LogLk(D/M) + Log(#sites) x #parameters

52

- PartitionFinder?

The full probabilistic model

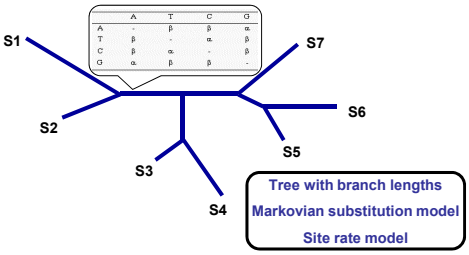


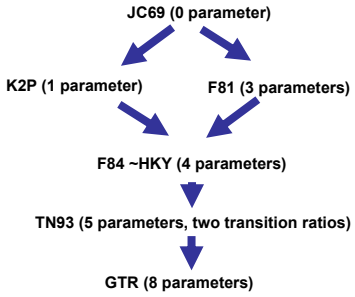
Tree with branch lengths

Markovian substitution model

Site rate model

The full probabilistic model





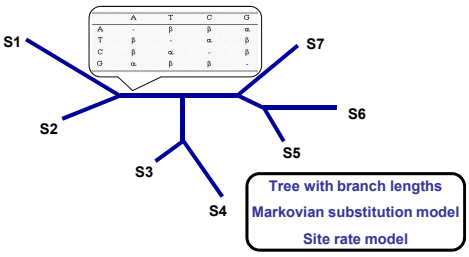
AIC= - 2 LogLk(D/M) + 2 x #parameters

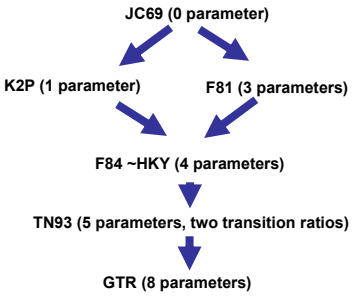
Log-likelihood GTR+I+G = -100

Log-likelihood HKY+G = -104

AIC value of each model ? Which one is best with AIC ?

The full probabilistic model





AIC= - 2 LogLk(D/M) + 2 x #parameters

Log-likelihood GTR+I+G = -100 AIC = 200 + 2 x (11 + 8 + 2) = 242

Log-likelihood HKY+G = -104 AIC = 208 + 2 x (11 + 4 + 1) = 240

AIC value of each model ? Which one is best with AIC ?

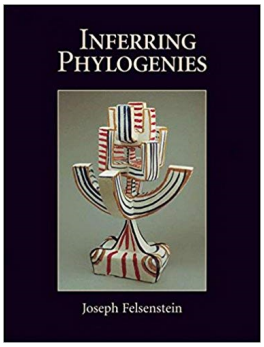
55

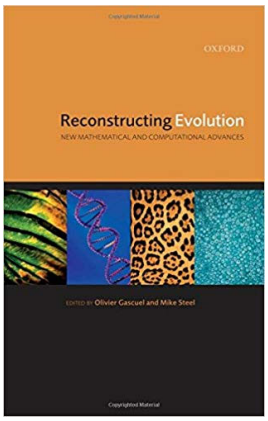
Model selection

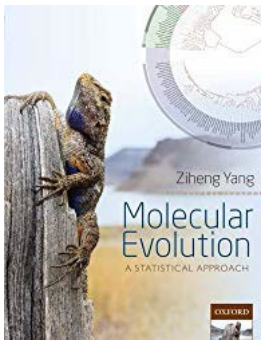
- The higher the likelihood of the data, the better the model
- But, with high number of parameters, we expect high likelihood values, because we have many degrees of freedom to fit the model to the data at hand.
- We then have to penalize the likelihood value to account for the number of parameters.
- In most cases we use AIC or BIC, to be minimized (thanks to software programs: ModelTest, ProtTest, SMS, ModelFinder…).
- Wrong again (sorry…), but super useful, especially with proteins, partition and mixture models, etc. (but DNA, GTR+I+G … ?)

56

References





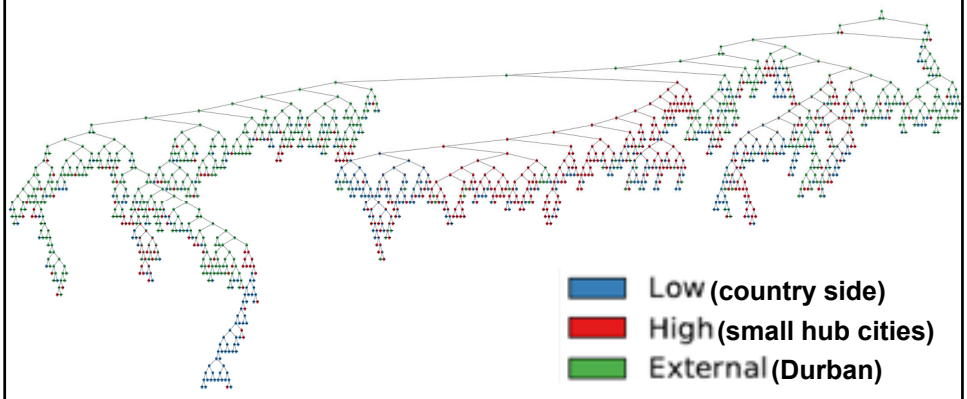


57

Ancestral state reconstructions

PastML MBE 2019, LSD Syst Biol 2015

Maximum likelihood tree and estimation of ancestral states, describing the phylodynamics of HIV infection in rural KwaZulu-Natal Global prevalence ~15%, much higher in young adults

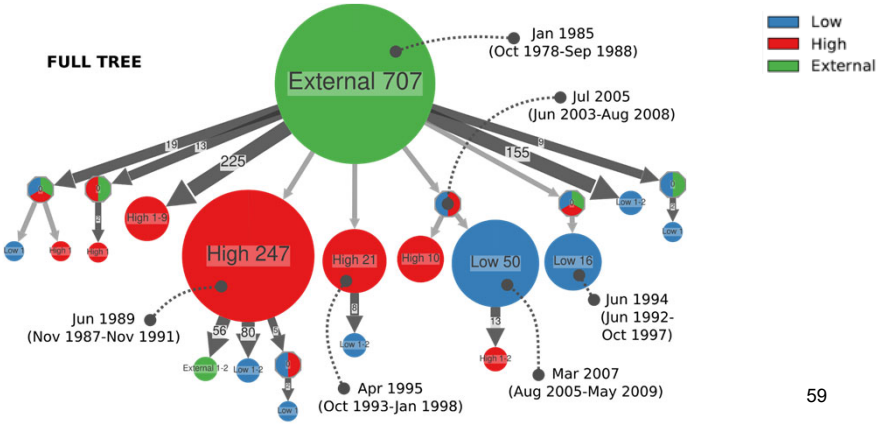


58

Ancestral state reconstructions

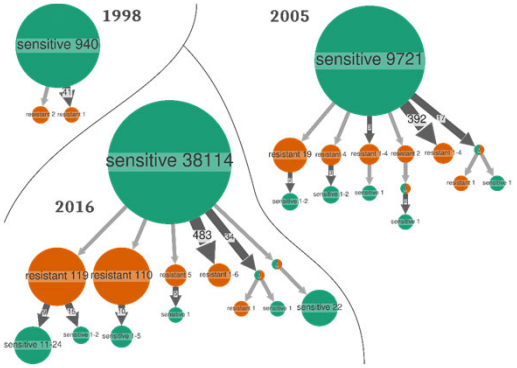
PastML MBE 2019, LSD Syst Biol 2015

Maximum likelihood tree and estimation of ancestral states, describing the phylodynamics of HIV infection in rural KwaZulu-Natal
Global prevalence ~15%, much higher in young adults



Ancestral state reconstructions

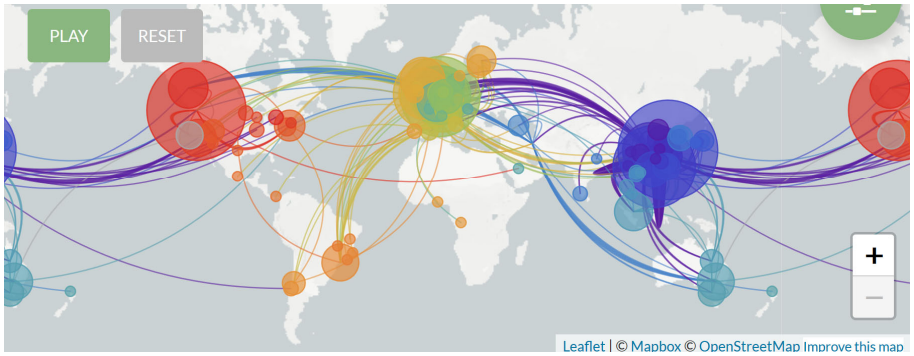
PastML MBE 2019, LSD Syst Biol 2015, Viruses 2023



UK – Subtype B - 40,000 sequences
DRM: L90M – Protease Inhibitors (introduced in 1996)
Medium fitness cost – Small resistance clusters
Preval. 5% -> 2% – 50/50 acquired/transmitted

Ancestral state reconstructions

NextStrain, March 2020



SARS-CoV-2 worldwide surveillance

61