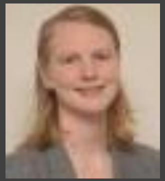
Dr. Gregory A. Babbitt

BABBITTLAB@RIT

COMPUTATIONAL MOLECULAR EVOLUTION AND ECOLOGY

some of our students

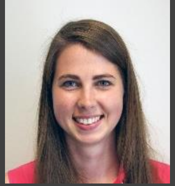
Erin Coppola BME



Jamie Mortensen BME

Justin Liao BME

Mohammed Alawad
Bioinformatics



Katharina Schulze
Bioinformatics

We are a mostly “dry-bench” or computer-based laboratory studying the evolution of the molecular components of cells. Biophysics and biochemistry underlies all molecular processes in the cell, but only some molecular structure and behavior, genetic and epigenetic, can contain information that is subject to heritable change over potentially deep timescales.

Our long-term research goals are aimed at a more biophysically-grounded understanding of molecular evolution in the cell. We combine biophysical and molecular evolutionary modeling to investigate the boundary between evolvable and non-evolvable biophysics.

Realistic expectations about genetic variants (especially human) requires an appreciation of how evolution and migration affects molecular evolution

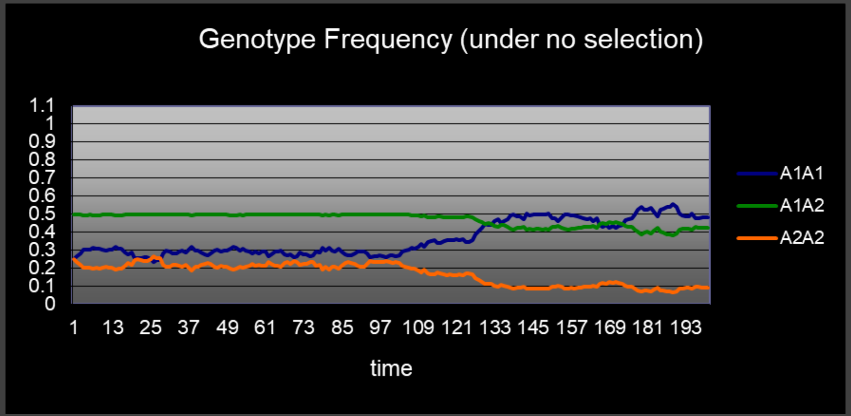
1. Almost all molecular variant evolution is neutral

2. Small populations randomly fix traits in populations

Wright-Fisher two allele dynamic pop gen model

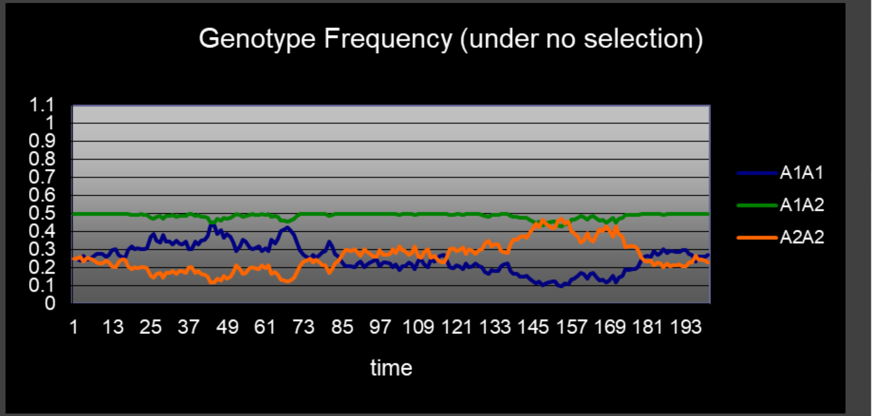
Pop size

2000



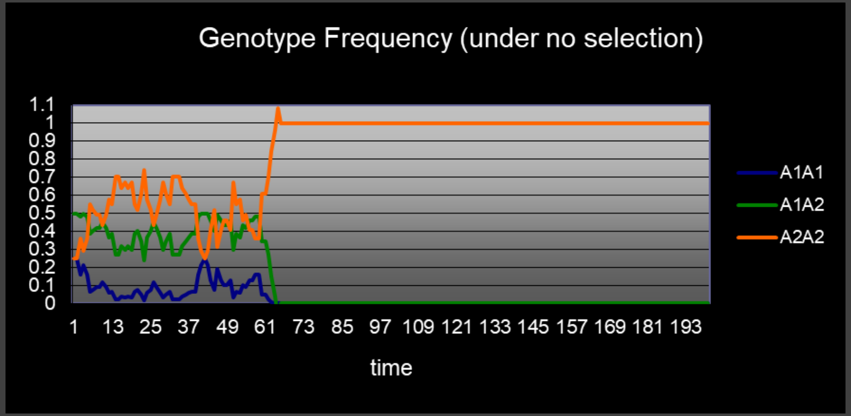
Genotype Frequency (under no selection)

500



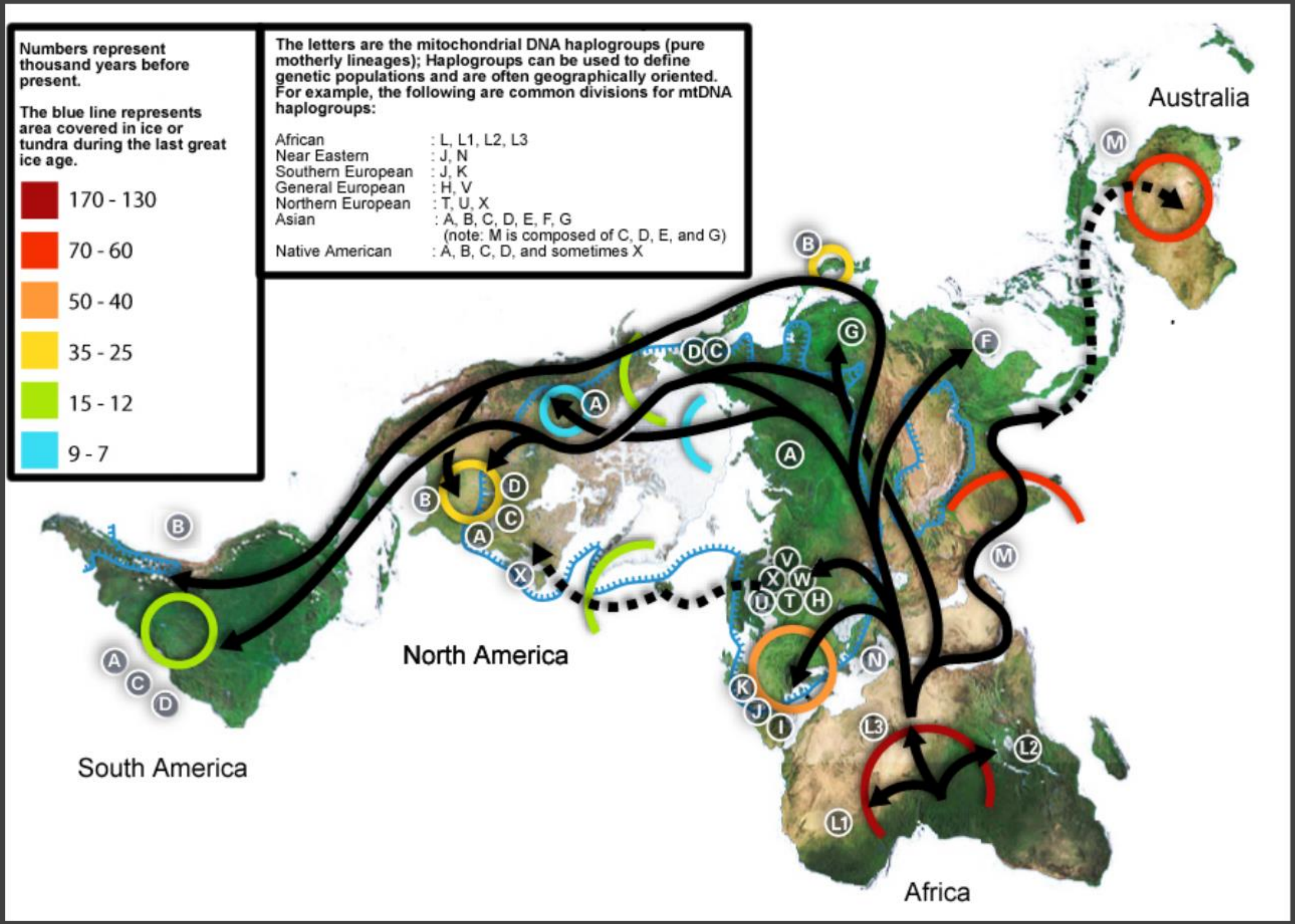
Genotype Frequency (under no selection)

50



Genotype Frequency (under no selection)

Human haplotype variation



Numbers represent thousand years before present.

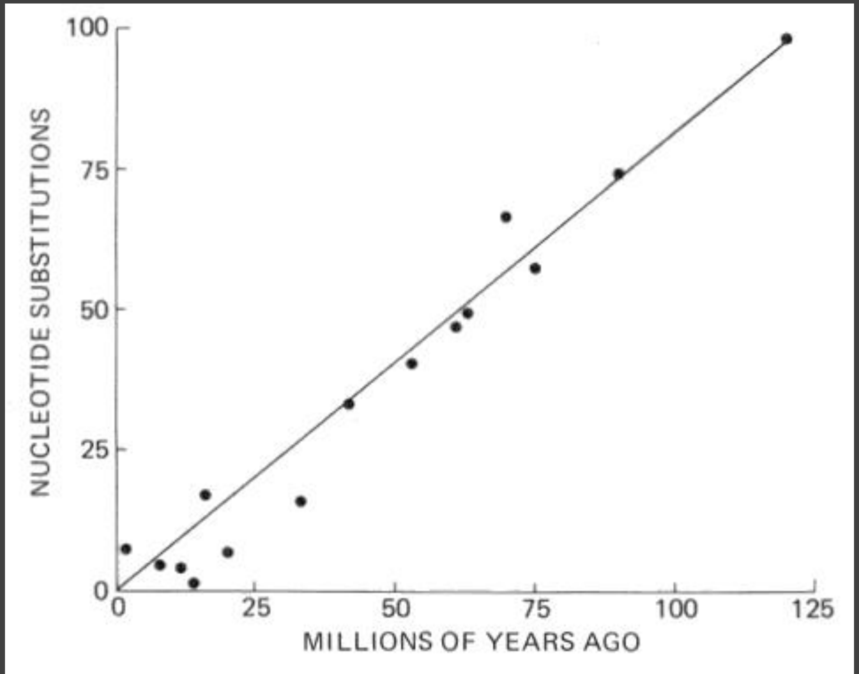
The blue line represents area covered in ice or tundra during the last great ice age.

🟥	170 - 130
🟥	70 - 60
🟧	50 - 40
🟨	35 - 25
🟩	15 - 12
🟦	9 - 7

The letters are the mitochondrial DNA haplogroups (pure motherly lineages); Haplogroups can be used to define genetic populations and are often geographically oriented. For example, the following are common divisions for mtDNA haplogroups:

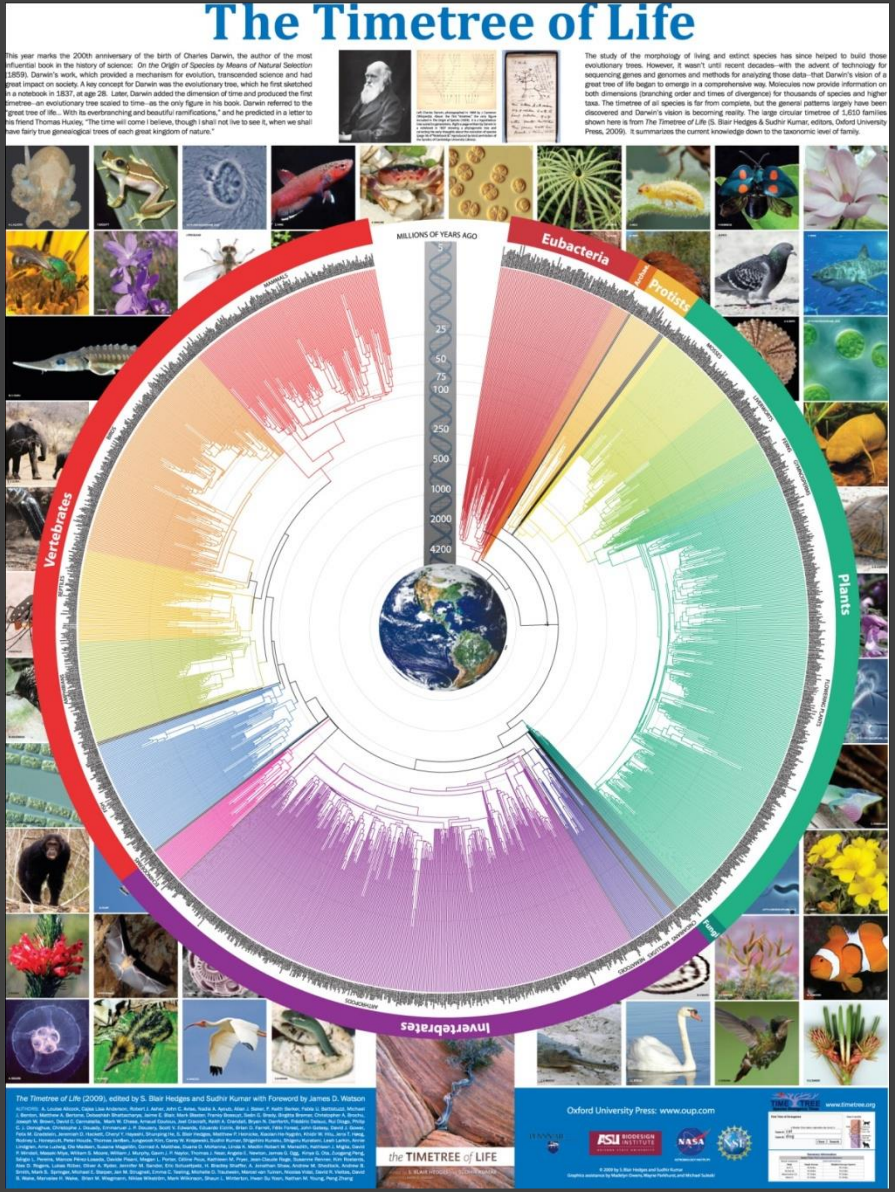
African	: L, L1, L2, L3
Near Eastern	: J, N
Southern European	: J, K
General European	: H, V
Northern European	: T, U, X
Asian	: A, B, C, D, E, F, G (note: M is composed of C, D, E, and G)
Native American	: A, B, C, D, and sometimes X

Phylogenetic reconstruction from aligned sequence data



Substitutions accumulate in a “clock-like” linear fashion ….. and so divergence times can be estimated by “molecular clocks”

…most mutations are assumed neutral in their effects …”neutral theory of mol evol”
Many are synonymous and/or fixed randomly



This year marks the 200th anniversary of the birth of Charles Darwin, the author of the most influential book in the history of science: *On the Origin of Species by Means of Natural Selection* (1859). Darwin's work, which provided a mechanism for evolution, transcended science and had great impact on society. A key concept for Darwin was the evolutionary tree, which he first sketched in a notebook in 1837, at age 28. Later, Darwin added the dimension of time and produced the first timetree--an evolutionary tree scaled to time--as the only figure in his book. Darwin referred to the "great tree of life… With its everbranching and beautiful ramifications," and he predicted in a letter to his friend Thomas Huxley, "The time will come I believe, though I shall not live to see it, when we shall have fairly true genealogical trees of each great kingdom of nature."

The study of the morphology of living and extinct species has since helped to build those evolutionary trees. However, it wasn't until recent decades--with the advent of technology for sequencing genes and genomes and methods for analyzing those data--that Darwin's vision of a great tree of life began to emerge in a comprehensive way. Molecules now provide information on both dimensions (branching order and times of divergence) for thousands of species and higher taxa. The timetree of all species is far from complete, but the general patterns largely have been discovered and Darwin's vision is becoming reality. The large circular timetree of 1,610 families shown here is from *The Timetree of Life* (S. Blair Hedges & Sudhir Kumar, editors, Oxford University Press, 2009). It summarizes the current knowledge down to the taxonomic level of family.

The Timetree of Life (2009), edited by S. Blair Hedges and Sudhir Kumar with Foreword by James D. Watson

AUTHORS: A. Louise Alcock, Cajsa Lisa Anderson, Robert J. Asher, John C. Avise, Nadia A. Ayoub, Allan J. Baker, F. Keith Barker, Fabia U. Battistuzzi, Michael J. Benton, Matthew A. Bertone, Debashish Bhattacharya, Jaime E. Blair, Mark Blaxter, Franky Bossuyt, Seán G. Brady, Birgitta Bremer, Christopher A. Brochu, Joseph W. Brown, David C. Cannatella, Mark W. Chase, Arnaud Couloux, Joel Cracraft, Keith A. Crandall, Bryan N. Danforth, Frédéric Delsuc, Rui Diogo, Philip C. J. Donoghue, Christophe J. Douady, Emmanuel J. P. Douzery, Scott V. Edwards, Eduardo Eizirik, Brian D. Farrell, Félix Forest, John Gatesy, David J. Gower, Felix M. Gradstein, Jeremiah D. Hackett, Cheryl Y. Hayashi, Shunping He, S. Blair Hedges, Matthew P. Heinicke, Xiaolan He-Nygrén, Khidir W. Hilu, Jens T. Høeg, Rodney L. Honeycutt, Peter Houde, Thomas Janßen, Jungwook Kim, Carey W. Krajewski, Sudhir Kumar, Shigehiro Kuraku, Shigeru Kuratani, Leah Larkin, Annie Lindgren, Arne Ludwig, Ole Madsen, Susana Magallón, Conrad A. Matthee, Duane D. McKenna, Linda K. Medlin, Robert W. Meredith, Kathleen J. Miglia, David P. Mindell, Masaki Miya, William S. Moore, William J. Murphy, Gavin J. P. Naylor, Thomas J. Near, Angela E. Newton, James G. Ogg, Kinya G. Ota, Zuogang Peng, Sérgio L. Pereira, Marcos Pérez-Losada, Davide Pisani, Megan L. Porter, Céline Poux, Kathleen M. Pryer, Jean-Claude Rage, Susanne Renner, Kim Roelants, Alex D. Rogers, Lukas Rüber, Oliver A. Ryder, Jennifer M. Sander, Eric Schuettpelz, H. Bradley Shaffer, A. Jonathan Shaw, Andrew M. Shedlock, Andrew B. Smith, Mark S. Springer, Michael E. Steiper, Jan M. Strugnell, Emma C. Teeling, Michelle D. Trautwein, Marcel van Tuinen, Nicolas Vidal, David R. Vieites, David B. Wake, Marvalee H. Wake, Brian M. Wiegmann, Niklas Wikström, Mark Wilkinson, Shaun L. Winterton, Hwan Su Yoon, Nathan M. Young, Peng Zhang

Oxford University Press: www.oup.com

www.timetree.org

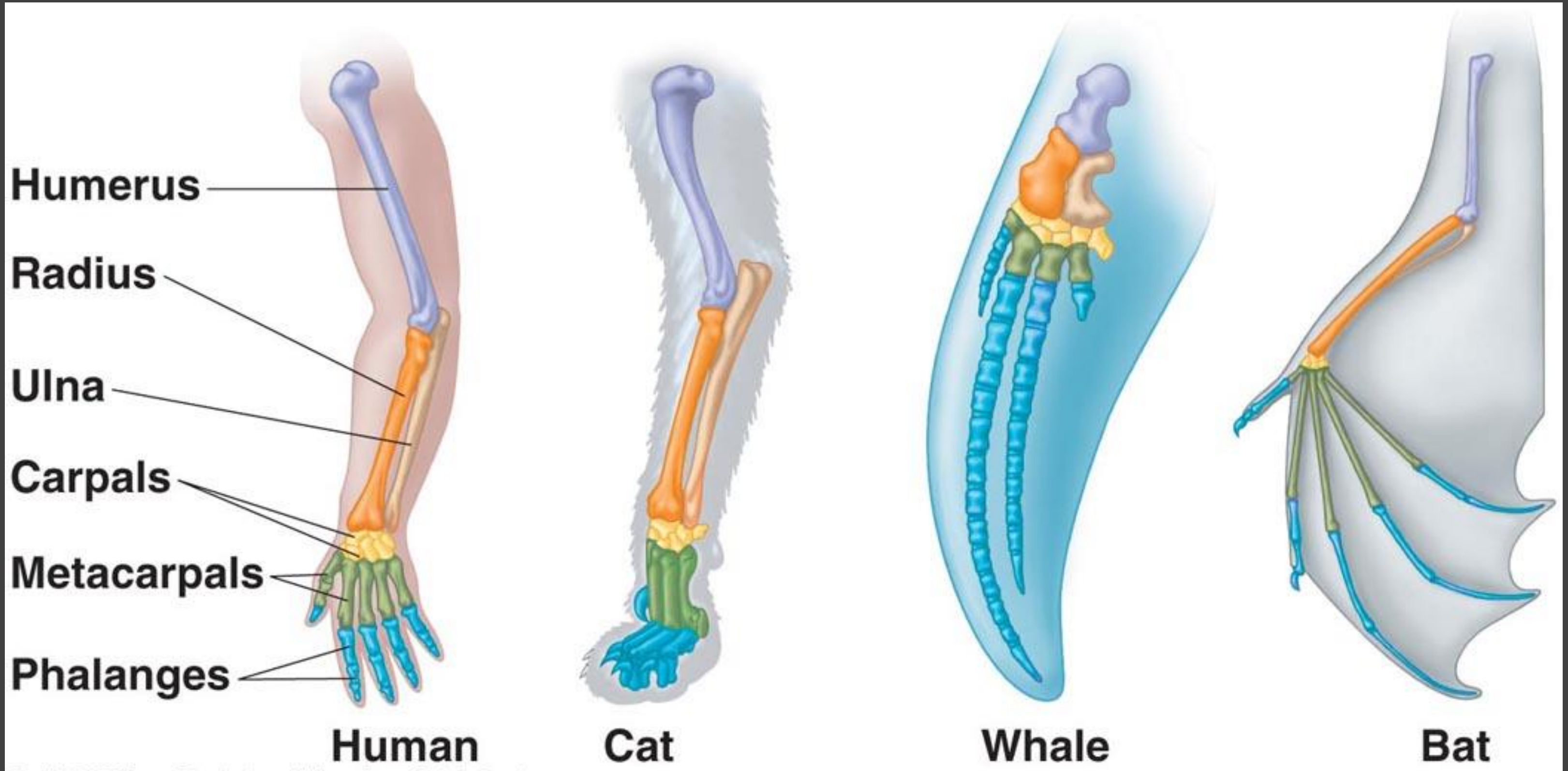
© 2009 by S. Blair Hedges and Sudhir Kumar
Graphics assistance by Madelyn Owens, Wayne Parkhurst, and Michael Suleski

Correct molecular homology is crucial for comparative genomics

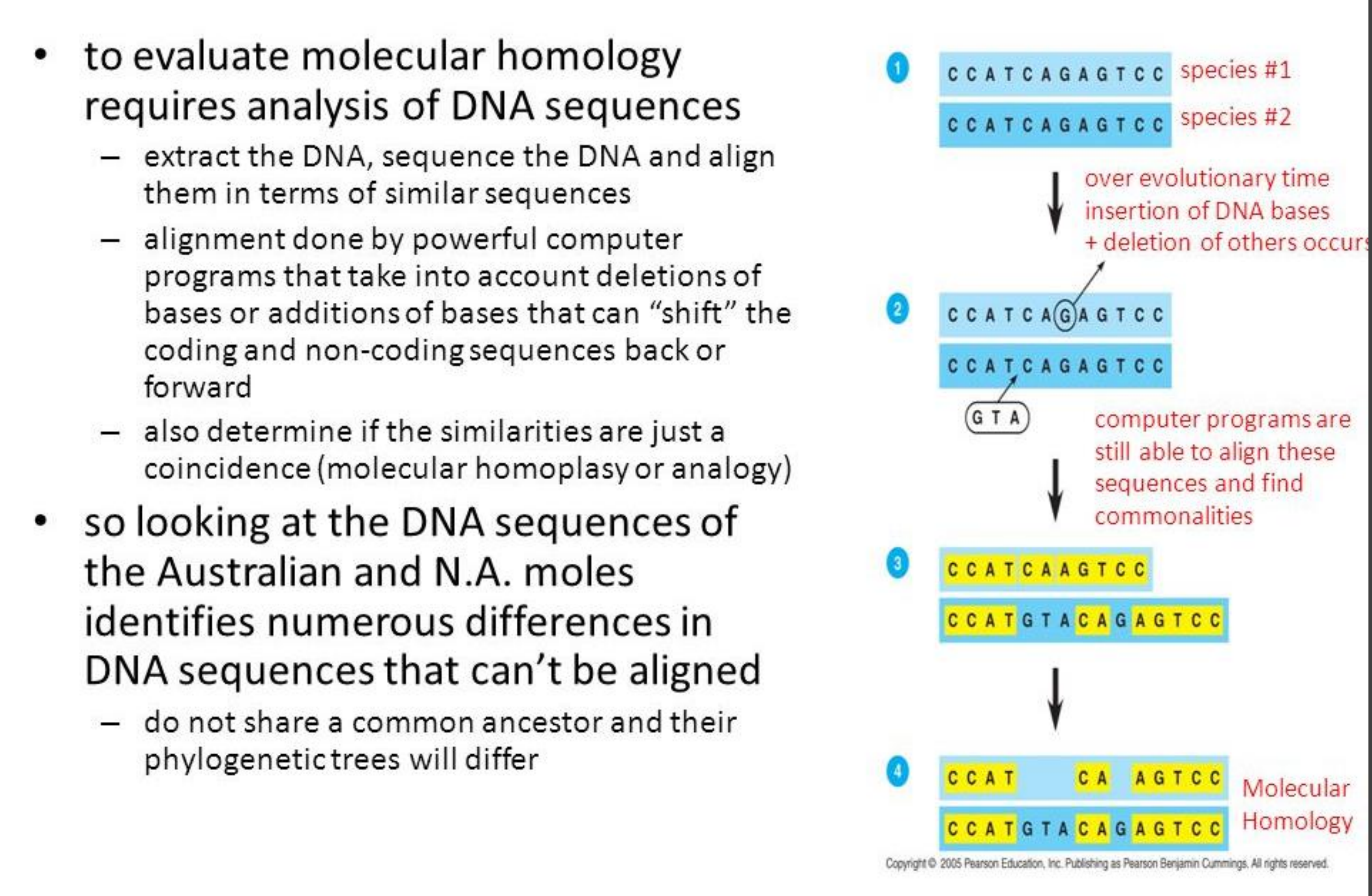
1. Scrutinize all results of alignment algorithms every time you use one (visually or using mismatch filter to mask poorly aligned regions)

2. Do not mix mutation events derived via gene duplication with speciation into same phylogenetic inference (i.e. analyze orthologs and paralogs separately)

character homology – paleontology



molecular homology – sequence level



- to evaluate molecular homology requires analysis of DNA sequences
  - extract the DNA, sequence the DNA and align them in terms of similar sequences
  - alignment done by powerful computer programs that take into account deletions of bases or additions of bases that can "shift" the coding and non-coding sequences back or forward
  - also determine if the similarities are just a coincidence (molecular homoplasy or analogy)
- so looking at the DNA sequences of the Australian and N.A. moles identifies numerous differences in DNA sequences that can't be aligned
  - do not share a common ancestor and their phylogenetic trees will differ

species #1

species #2

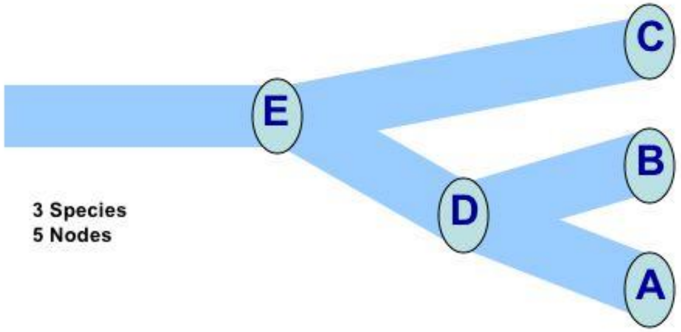
over evolutionary time insertion of DNA bases + deletion of others occurs

computer programs are still able to align these sequences and find commonalities

Molecular Homology

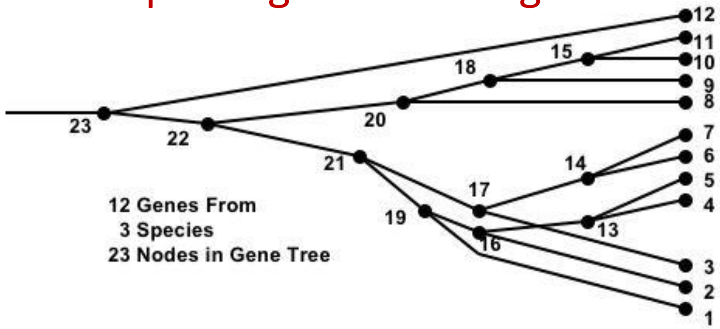
Species Tree

orthologs

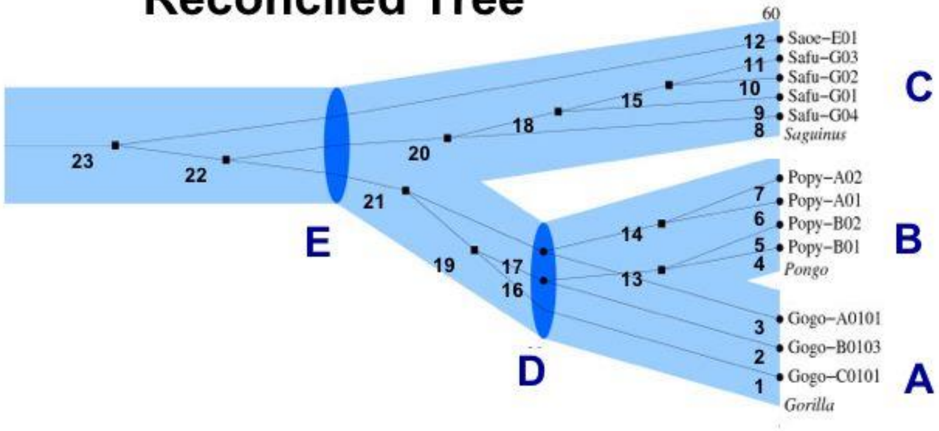


Gene Tree

paralogs = “Ohnologs”

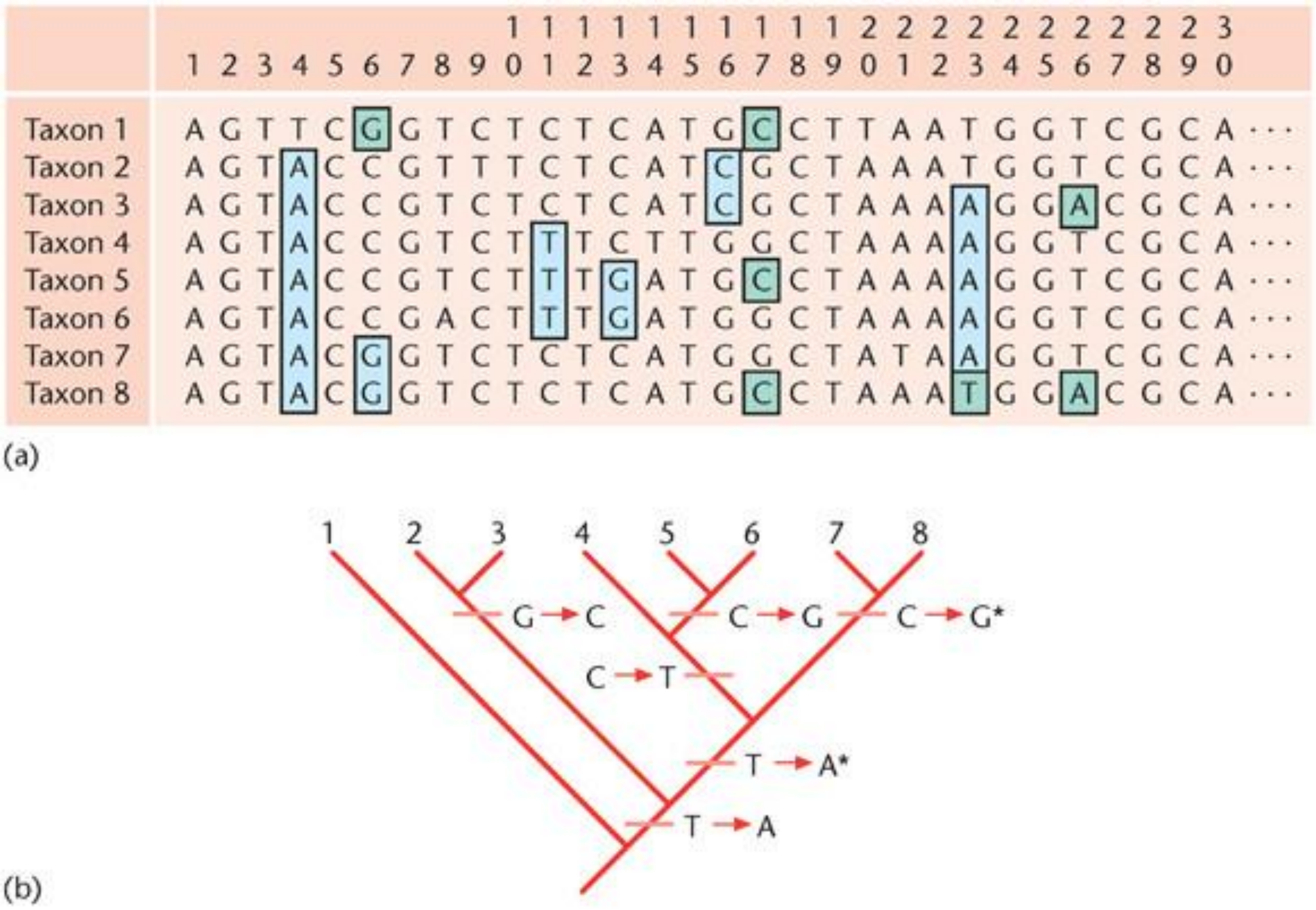


Reconciled Tree



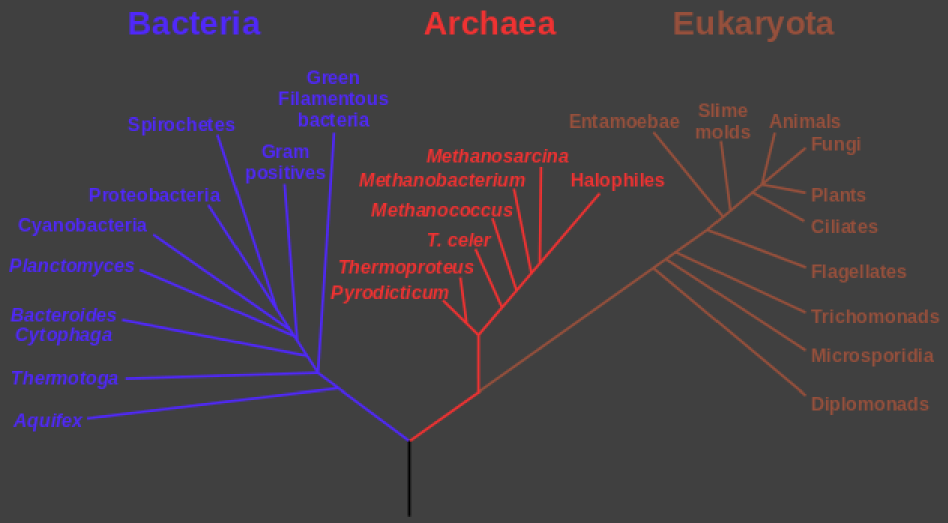
Inferring the most probable tree topology

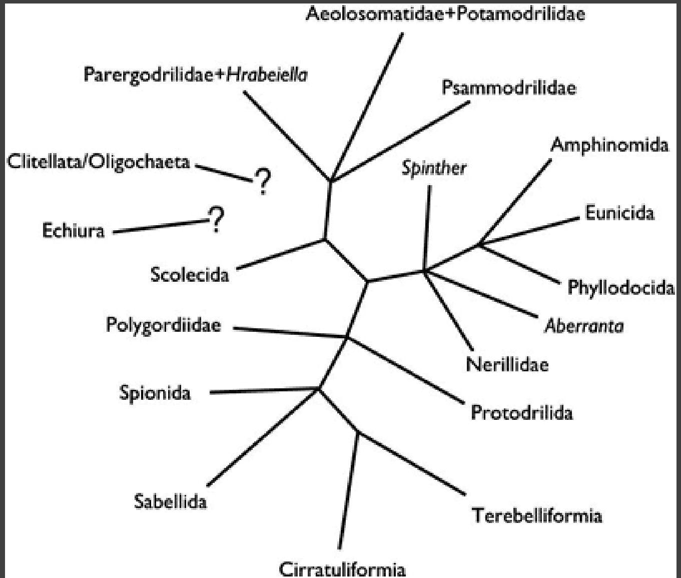
phylogenetic inference

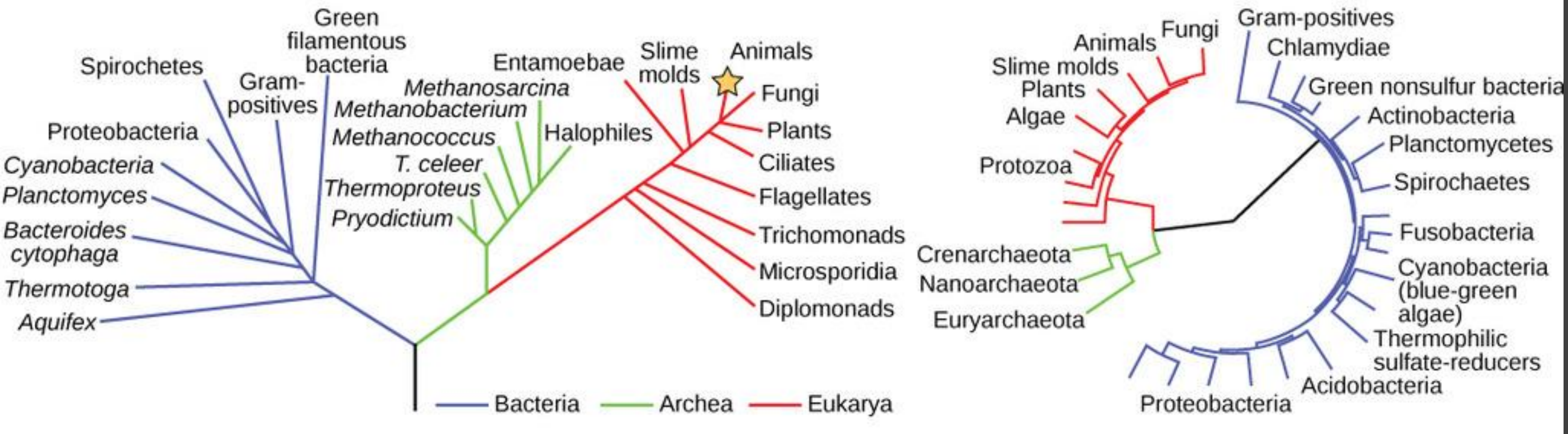


										1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	3	
	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	
Taxon 1	A	G	T	T	C	G	G	T	C	T	C	T	C	A	T	G	C	C	T	T	A	A	T	G	G	T	C	G	C	A	···
Taxon 2	A	G	T	A	C	C	G	T	T	T	C	T	C	A	T	C	G	C	T	A	A	A	T	G	G	T	C	G	C	A	···
Taxon 3	A	G	T	A	C	C	G	T	C	T	C	T	C	A	T	C	G	C	T	A	A	A	A	G	G	A	C	G	C	A	···
Taxon 4	A	G	T	A	C	C	G	T	C	T	T	T	C	T	T	G	G	C	T	A	A	A	A	G	G	T	C	G	C	A	···
Taxon 5	A	G	T	A	C	C	G	T	C	T	T	T	G	A	T	G	C	C	T	A	A	A	A	G	G	T	C	G	C	A	···
Taxon 6	A	G	T	A	C	C	G	A	C	T	T	T	G	A	T	G	G	C	T	A	A	A	A	G	G	T	C	G	C	A	···
Taxon 7	A	G	T	A	C	G	G	T	C	T	C	T	C	A	T	G	G	C	T	A	T	A	A	G	G	T	C	G	C	A	···
Taxon 8	A	G	T	A	C	G	G	T	C	T	C	T	C	A	T	G	C	C	T	A	A	A	T	G	G	A	C	G	C	A	···

phylogenetic tree viewing



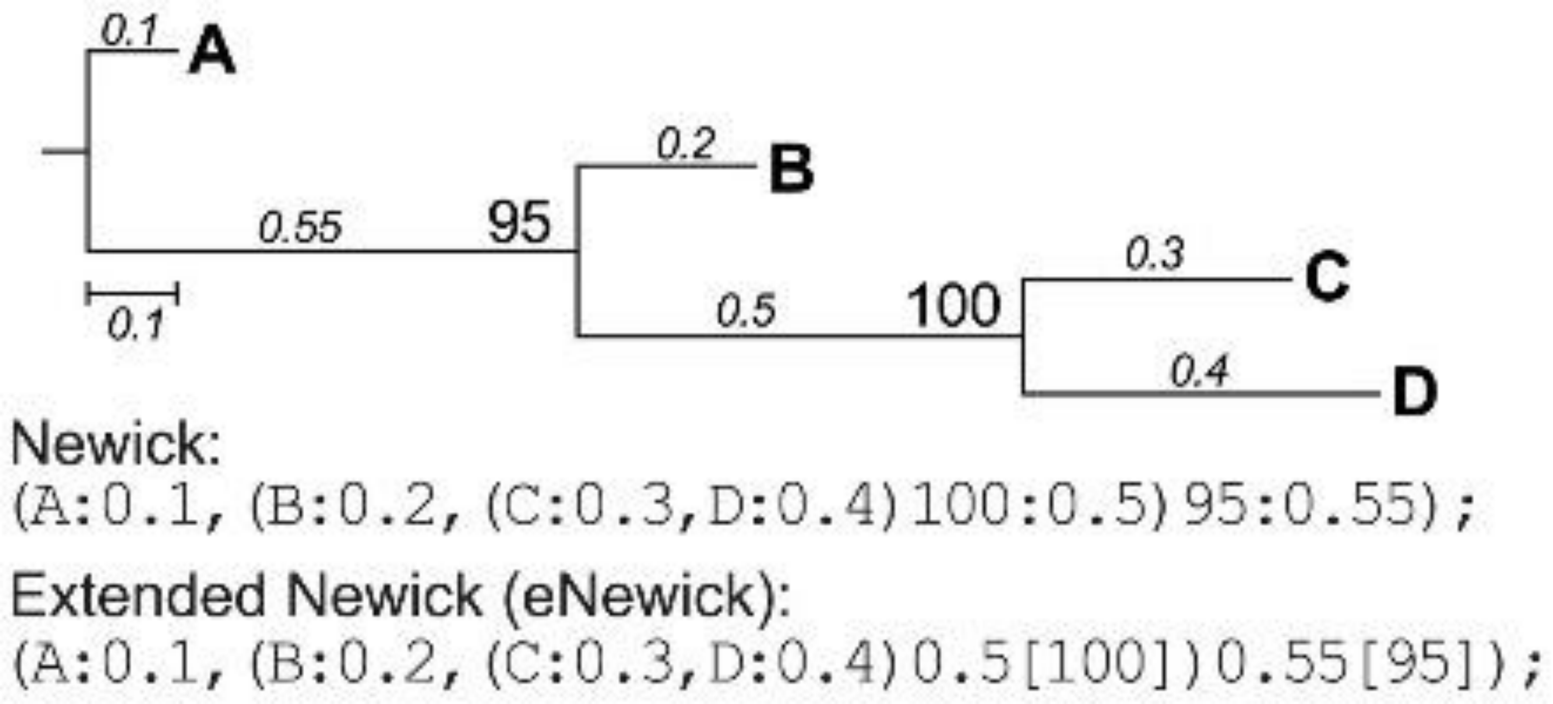




(a) Rooted phylogenetic tree

(b) Unrooted phylogenetic tree

phylogenetic tree representation – computer format



Newick:

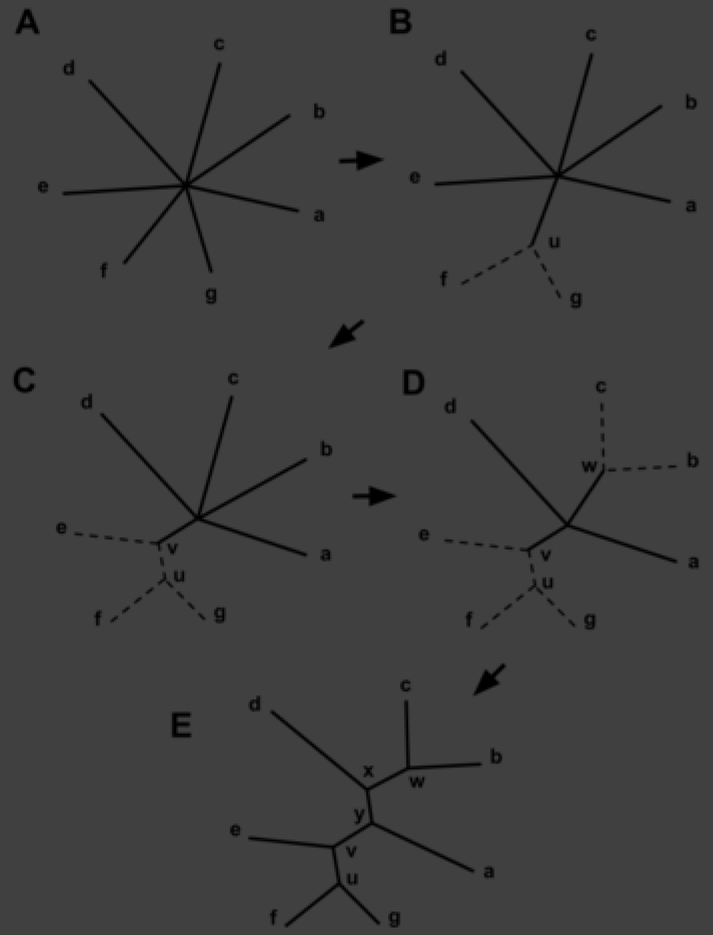
```
(A:0.1,(B:0.2,(C:0.3,D:0.4)100:0.5)95:0.55);
```

Extended Newick (eNewick):

```
(A:0.1,(B:0.2,(C:0.3,D:0.4)0.5[100])0.55[95]);
```

Distance-based phylogenetic reconstruction from aligned sequence data (neighbor-joining algorithm)

Starting with a star tree (A), the Q matrix is calculated and used to choose a pair of nodes for joining, in this case f and g. These are joined to a newly created node, u, as shown in (B). The part of the tree shown as dotted lines is now fixed and will not be changed in subsequent joining steps. The distances from node u to the nodes a-e are computed from the formula given in the text. This process is then repeated, using a matrix of just the distances between the nodes, a,b,c,d,e, and u, and a Q matrix derived from it. In this case u and e are joined to the newly created v, as shown in (C). Two more iterations lead first to (D), and then to (E), at which point the algorithm is done, as the tree is fully resolved



Based on a distance matrix relating the r taxa, calculate Q as follows:

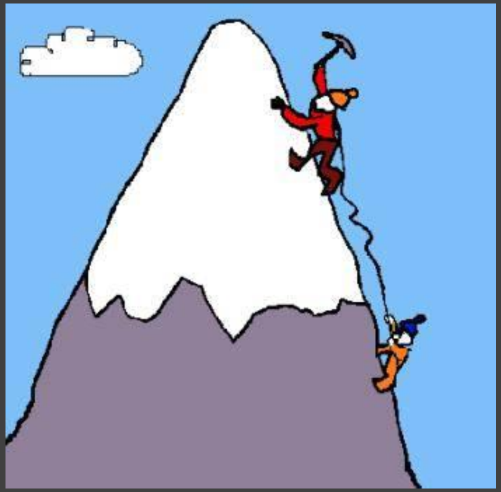
$$Q(i,j) = (r-2)d(i,j) - \sum_{k=1}^{r} d(i,k) - \sum_{k=1}^{r} d(j,k)$$

where $d(i,j)$ is the distance between taxa i and j.

Note: evolutionary distances are just mutation counts adjusted for the probability of multiple hits at the same sites

Frequentist paradigm

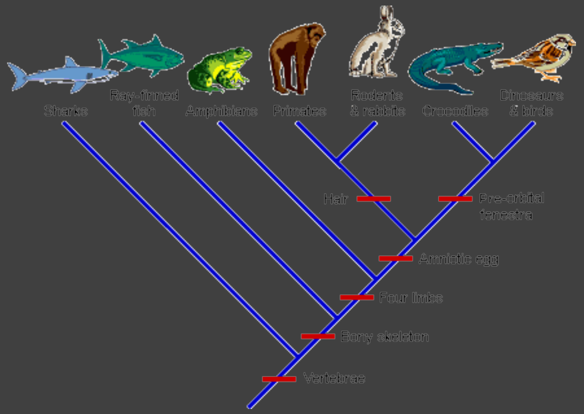
Maximum likelihood estimate



$$P(x) = \text{freq}\ (x \mid \theta)$$

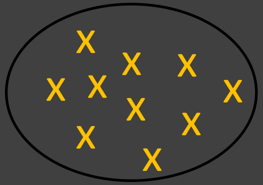
$$P(E) = \text{freq}\ (E \mid H)$$

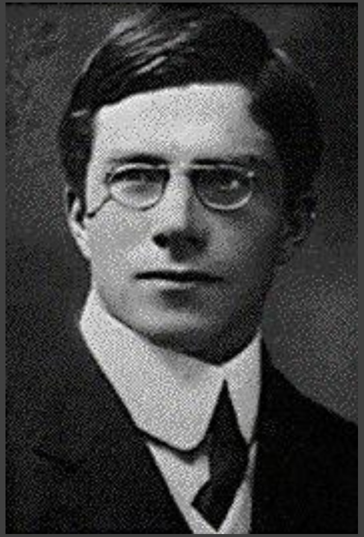
Maximum likelihood tree



Likelihoods are much more useful

$$L(H) = \text{prob}\ (H \mid E)$$

L(model) = joint prob (model | )



Ronald Alymer Fisher

Probability versus Likelihood

P(x) = P(occurrence of observation given particular model of known parameters)

More simply ….**P(data|model parameters)** Is this often useful?

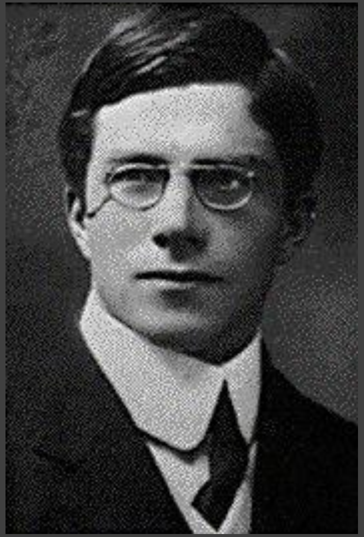
Do we always ‘know’ what model fits?

likelihood flips this relation **L(model parameters | data)**

Thus $L(\theta/x) = P(x/\theta)$

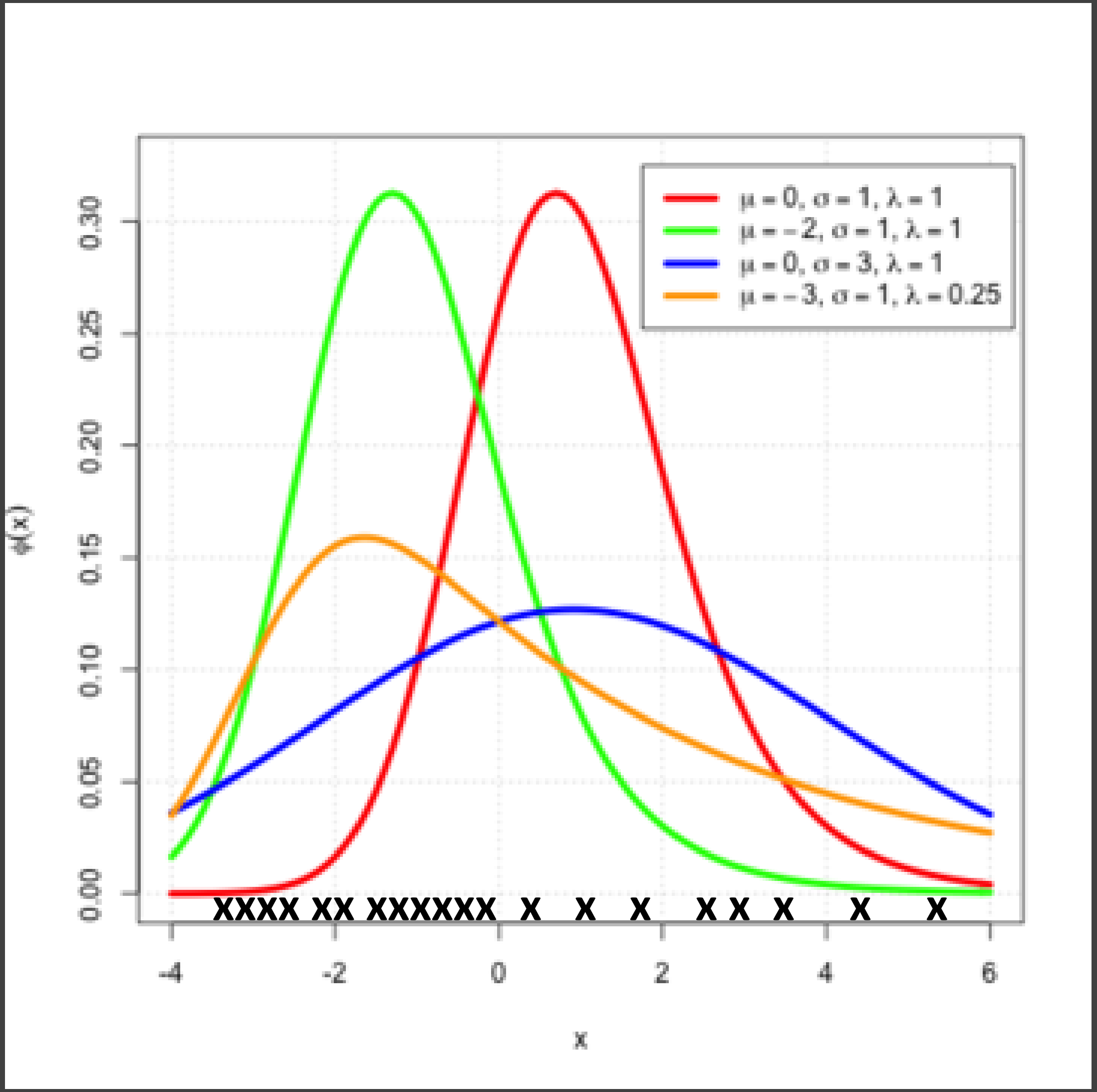
Now we can think of the likelihood of a model given a data point
….or better many data points

Thus $L(\theta; x_1 \ldots . x_n) = \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} ln f(x_1|\theta)$



Ronald Alymer Fisher

maximum likelihood estimation and multimodel inference



Ronald Fisher in 1913

Ronald Alymer Fisher

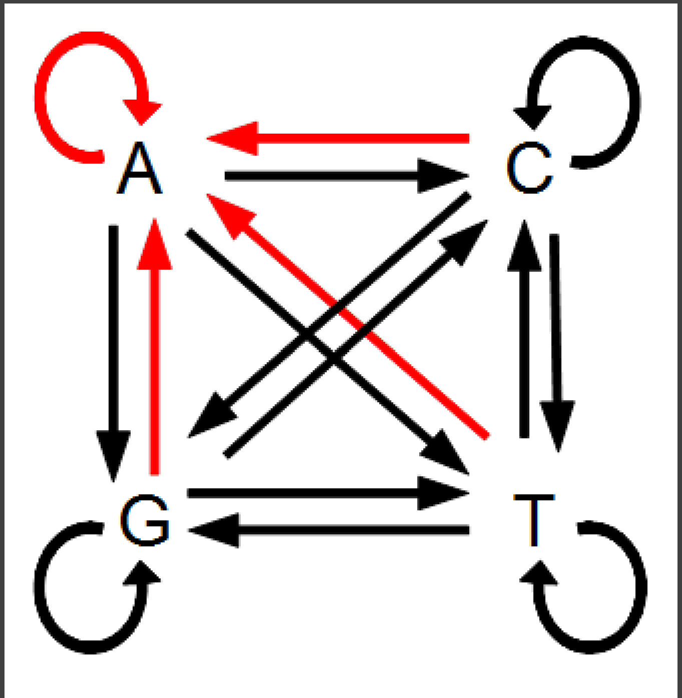
What is conditional probability is generally better to know?

P(model|data) or P(data|model)

Which density curve is most likely given the data below?

Underlying models of evolution (DNA substitution vs codon transitions)

Simplest model (Jukes Cantor JC69) all substitution
rates are assumed of equal potential probability
and related to overall mutation rate (i.e. one parameter model)



$$Q = \begin{pmatrix} -\mu_A & \mu_{GA} & \mu_{CA} & \mu_{TA} \\ \mu_{AG} & -\mu_G & \mu_{CG} & \mu_{TG} \\ \mu_{AC} & \mu_{GC} & -\mu_C & \mu_{TC} \\ \mu_{AT} & \mu_{GT} & \mu_{CT} & -\mu_T \end{pmatrix}$$

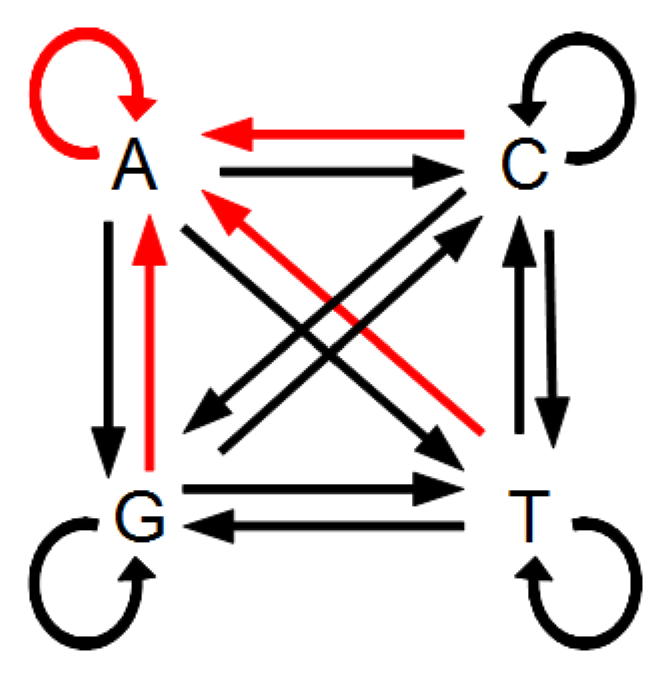
Subs rate matrix

$$P(t) = \begin{pmatrix} p_{AA}(t) & p_{GA}(t) & p_{CA}(t) & p_{TA}(t) \\ p_{AG}(t) & p_{GG}(t) & p_{CG}(t) & p_{TG}(t) \\ p_{AC}(t) & p_{GC}(t) & p_{CC}(t) & p_{TC}(t) \\ p_{AT}(t) & p_{GT}(t) & p_{CT}(t) & p_{TT}(t) \end{pmatrix}$$

probability matrix of observing subs – recall adjust for multiple hits

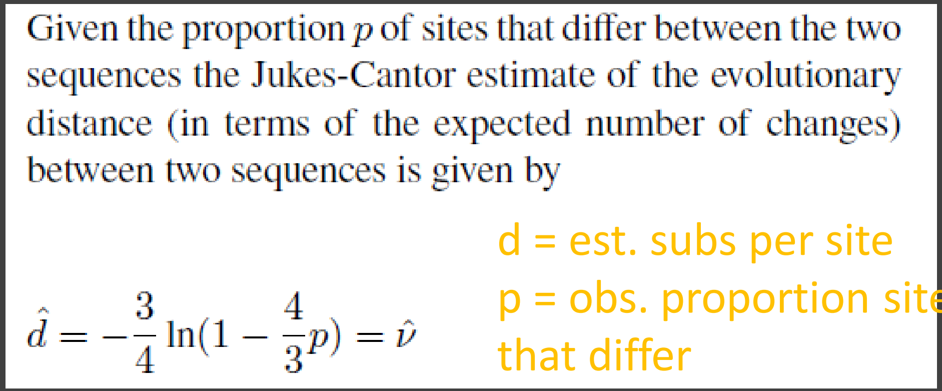
Underlying models of evolution (DNA substitution vs codon transitions)

Simplest model (Jukes Cantor JC69) all substitution
rates are assumed of equal potential probability
and related to overall mutation rate (i.e. one parameter model)



$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

Subs rate matrix



Given the proportion p of sites that differ between the two sequences the Jukes-Cantor estimate of the evolutionary distance (in terms of the expected number of changes) between two sequences is given by

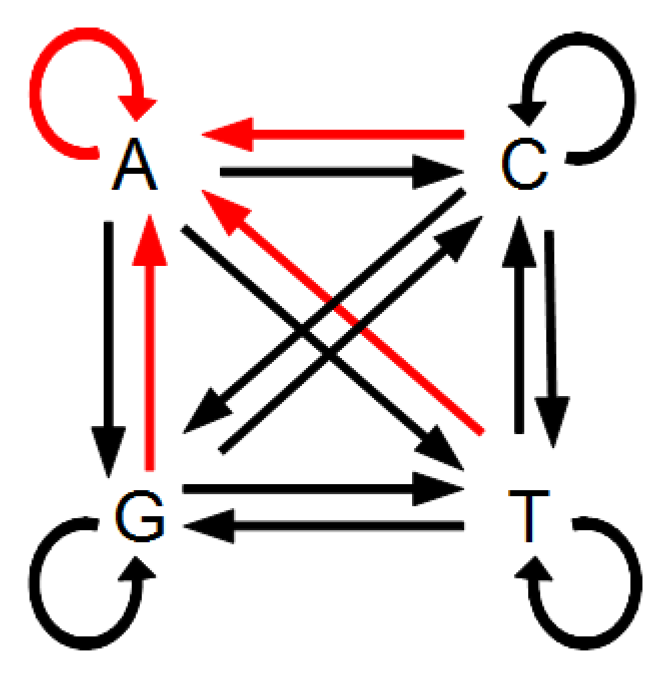
$$\hat{d} = -\frac{3}{4}\ln(1 - \frac{4}{3}p) = \hat{\nu}$$

d = est. subs per site
p = obs. proportion sites
that differ

probability matrix of observing subs – recall adjust for multiple hits

Underlying models of evolution (DNA substitution models)

Simplest model (Jukes Cantor JC69) all substitution
rates are assumed of equal potential probability
and related to overall mutation rate (i.e. one parameter model)



$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

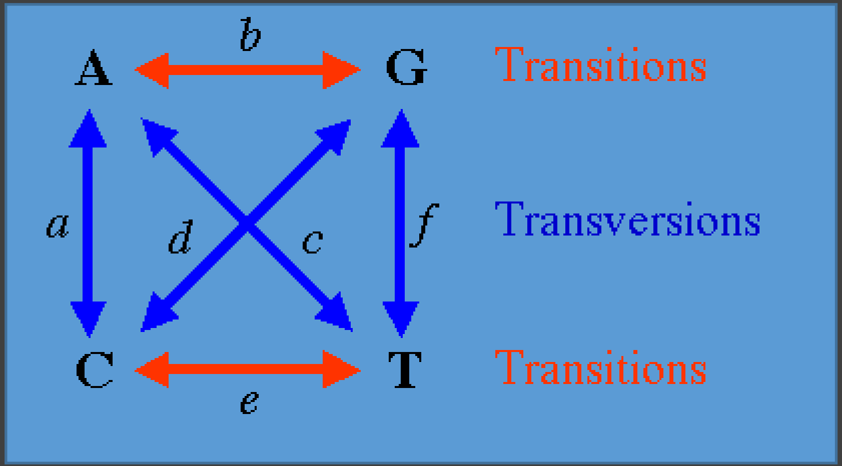
Subs rate matrix

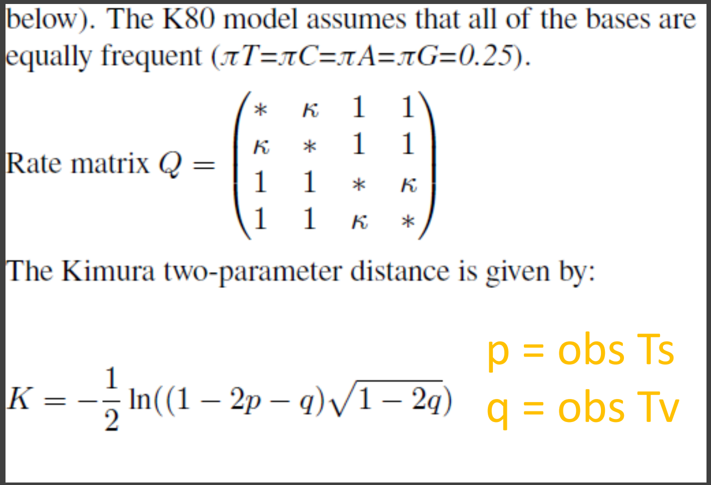
$$P = \begin{pmatrix} \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} \end{pmatrix}$$

probability matrix of observing subs – recall adjust for multiple hits

Underlying models of evolution (DNA substitution models)

K80 model – substitution rates can differ as two parameter model




below). The K80 model assumes that all of the bases are equally frequent ($\pi T = \pi C = \pi A = \pi G = 0.25$).

$$\text{Rate matrix } Q = \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix}$$

The Kimura two-parameter distance is given by:

$$K = -\frac{1}{2}\ln((1-2p-q)\sqrt{1-2q})$$

p = obs Ts
q = obs Tv

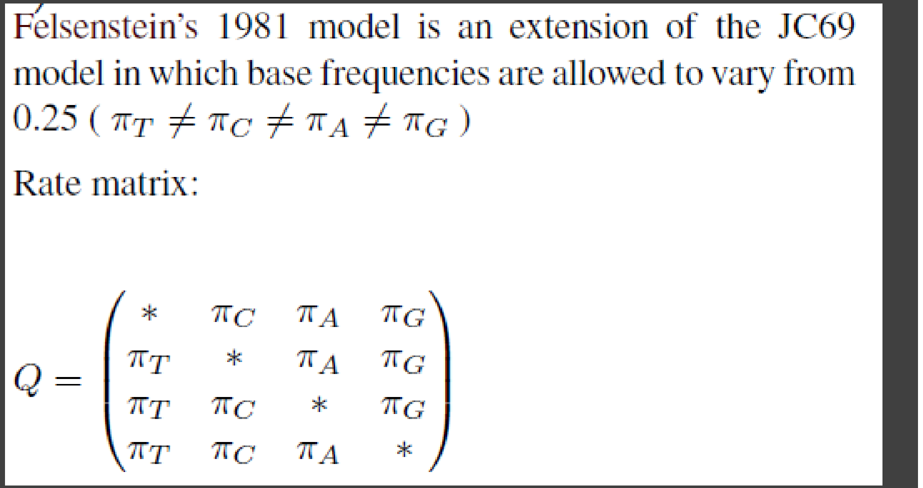
Ts = transitions – do NOT change purine-pyrimidine state of sequences
Tv = transversions – do change purine-pyrimidine state of sequences

Ts are more common than Tv despite having half as many ways to happen
…this is largely due to spontaneous deamination of C -> T over time

In truth…mutational spectrum observed in MA lines in model organisms never seem to match those inferred in comparative genomics. We (my lab) thinks this is due To universal weak selection on physical aspects of the genome (more to come…)

Underlying models of evolution (DNA substitution models)

F81 model – base composition can vary

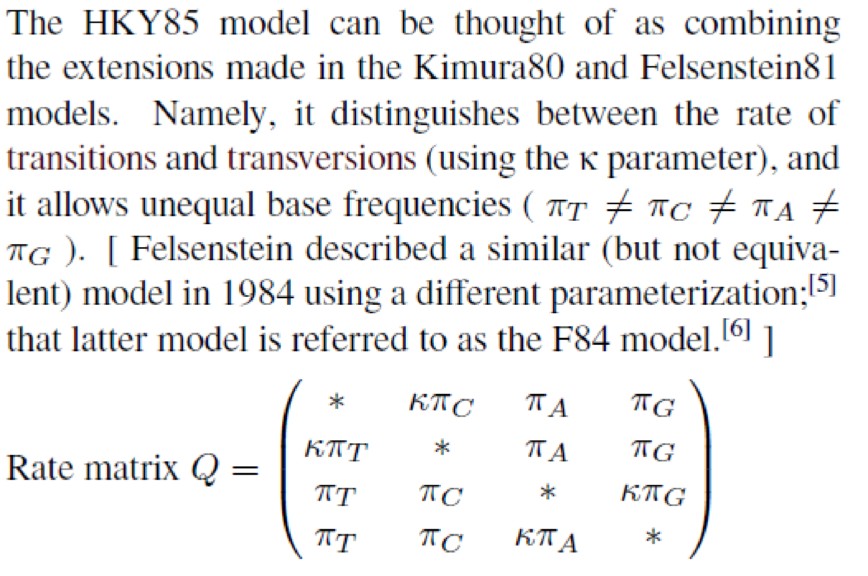


Felsenstein's 1981 model is an extension of the JC69 model in which base frequencies are allowed to vary from 0.25 ($\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$)

Rate matrix:

$$Q = \begin{pmatrix} * & \pi_C & \pi_A & \pi_G \\ \pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \pi_G \\ \pi_T & \pi_C & \pi_A & * \end{pmatrix}$$

HKY85 model – accomodates base comp and Ts:Tv ratio



The HKY85 model can be thought of as combining the extensions made in the Kimura80 and Felsenstein81 models. Namely, it distinguishes between the rate of transitions and transversions (using the κ parameter), and it allows unequal base frequencies ($\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$). [Felsenstein described a similar (but not equivalent) model in 1984 using a different parameterization;[5] that latter model is referred to as the F84 model.[6]]

$$\text{Rate matrix } Q = \begin{pmatrix} * & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & * \end{pmatrix}$$

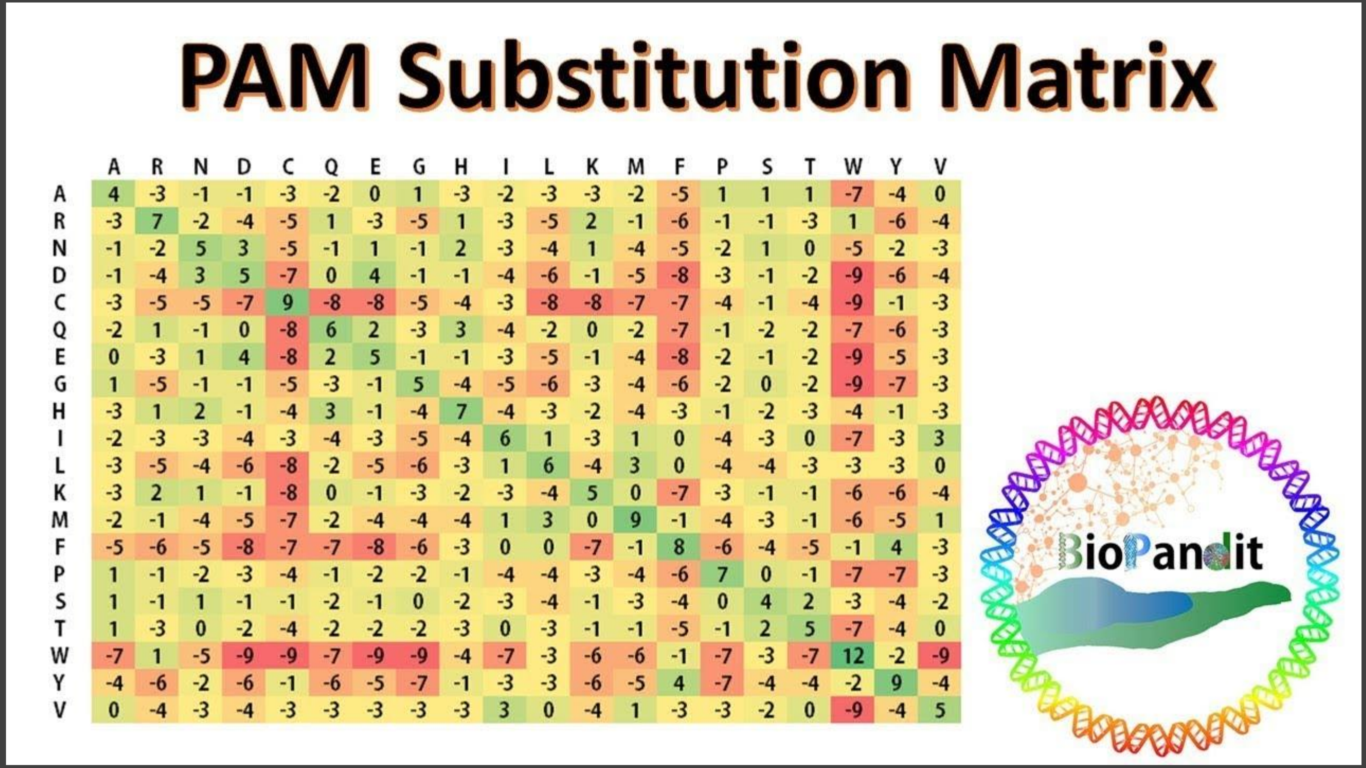
Other models

T92 – extends K80 to accommodate GC content change

TN93 – distinguishes between two types of Ts

GTR – general time reversible – 6 subs rate parameters and 4 equilibrium base
freq parameters (206 parameters for 20 aa's)
(633 parameters for all 64 codons)

Protein evolution models - Dayhoff, JTT, WAG and LG models



PAM Substitution Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-3	-1	-1	-3	-2	0	1	-3	-2	-3	-3	-2	-5	1	1	1	-7	-4	0
R	-3	7	-2	-4	-5	1	-3	-5	1	-3	-5	2	-1	-6	-1	-1	-3	1	-6	-4
N	-1	-2	5	3	-5	-1	1	-1	2	-3	-4	1	-4	-5	-2	1	0	-5	-2	-3
D	-1	-4	3	5	-7	0	4	-1	-1	-4	-6	-1	-5	-8	-3	-1	-2	-9	-6	-4
C	-3	-5	-5	-7	9	-8	-8	-5	-4	-3	-8	-8	-7	-7	-4	-1	-4	-9	-1	-3
Q	-2	1	-1	0	-8	6	2	-3	3	-4	-2	0	-2	-7	-1	-2	-2	-7	-6	-3
E	0	-3	1	4	-8	2	5	-1	-1	-3	-5	-1	-4	-8	-2	-1	-2	-9	-5	-3
G	1	-5	-1	-1	-5	-3	-1	5	-4	-5	-6	-3	-4	-6	-2	0	-2	-9	-7	-3
H	-3	1	2	-1	-4	3	-1	-4	7	-4	-3	-2	-4	-3	-1	-2	-3	-4	-1	-3
I	-2	-3	-3	-4	-3	-4	-3	-5	-4	6	1	-3	1	0	-4	-3	0	-7	-3	3
L	-3	-5	-4	-6	-8	-2	-5	-6	-3	1	6	-4	3	0	-4	-4	-3	-3	-3	0
K	-3	2	1	-1	-8	0	-1	-3	-2	-3	-4	5	0	-7	-3	-1	-1	-6	-6	-4
M	-2	-1	-4	-5	-7	-2	-4	-4	-4	1	3	0	9	-1	-4	-3	-1	-6	-5	1
F	-5	-6	-5	-8	-7	-7	-8	-6	-3	0	0	-7	-1	8	-6	-4	-5	-1	4	-3
P	1	-1	-2	-3	-4	-1	-2	-2	-1	-4	-4	-3	-4	-6	7	0	-1	-7	-7	-3
S	1	-1	1	-1	-1	-2	-1	0	-2	-3	-4	-1	-3	-4	0	4	2	-3	-4	-2
T	1	-3	0	-2	-4	-2	-2	-2	-3	0	-3	-1	-1	-5	-1	2	5	-7	-4	0
W	-7	1	-5	-9	-9	-7	-9	-9	-4	-7	-3	-6	-6	-1	-7	-3	-7	12	-2	-9
Y	-4	-6	-2	-6	-1	-6	-5	-7	-1	-3	-3	-6	-5	4	-7	-4	-4	-2	9	-4
V	0	-4	-3	-4	-3	-3	-3	-3	-3	3	0	-4	1	-3	-3	-2	0	-9	-4	5

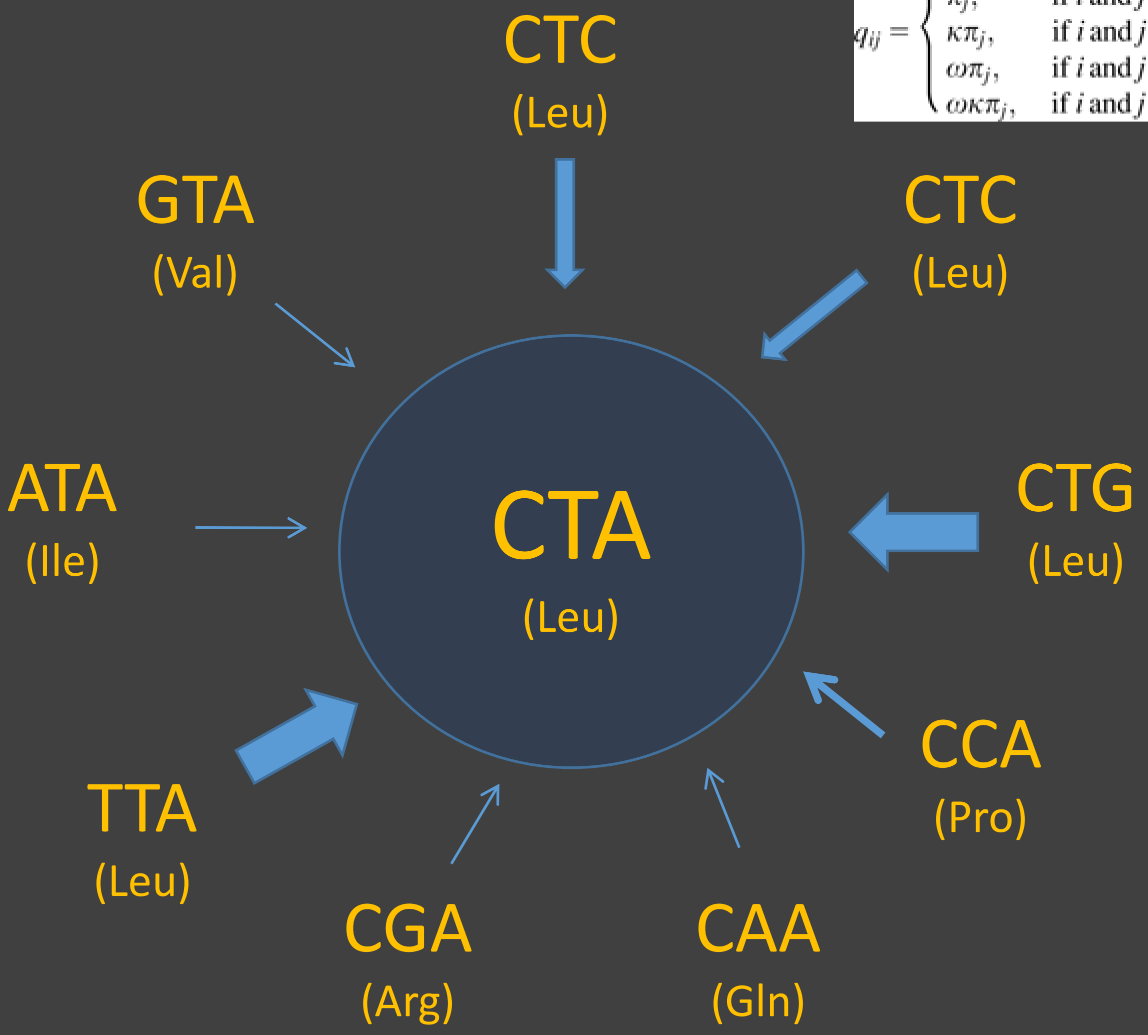
JTT – like PAM but based on more data

WAG – Whelan and Goldman
uses ML fitting very massive database to create subs matrix

LG – Le and Goldman
uses ML fitting of your data to create subs matrix

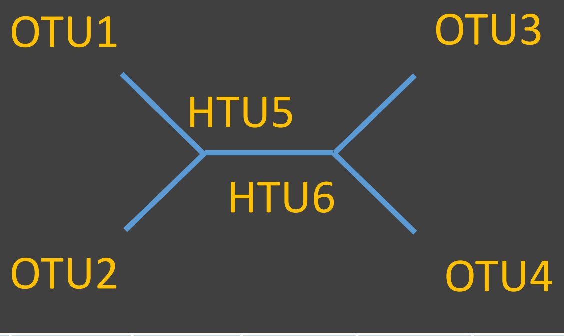
Codon evolution models - 64 X 64 matrix

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three positions,} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition,} \end{cases}$$



Purine = R
Pyrimidine = Y
Transition = R->R or Y->Y
Transversion = R->Y or Y->R

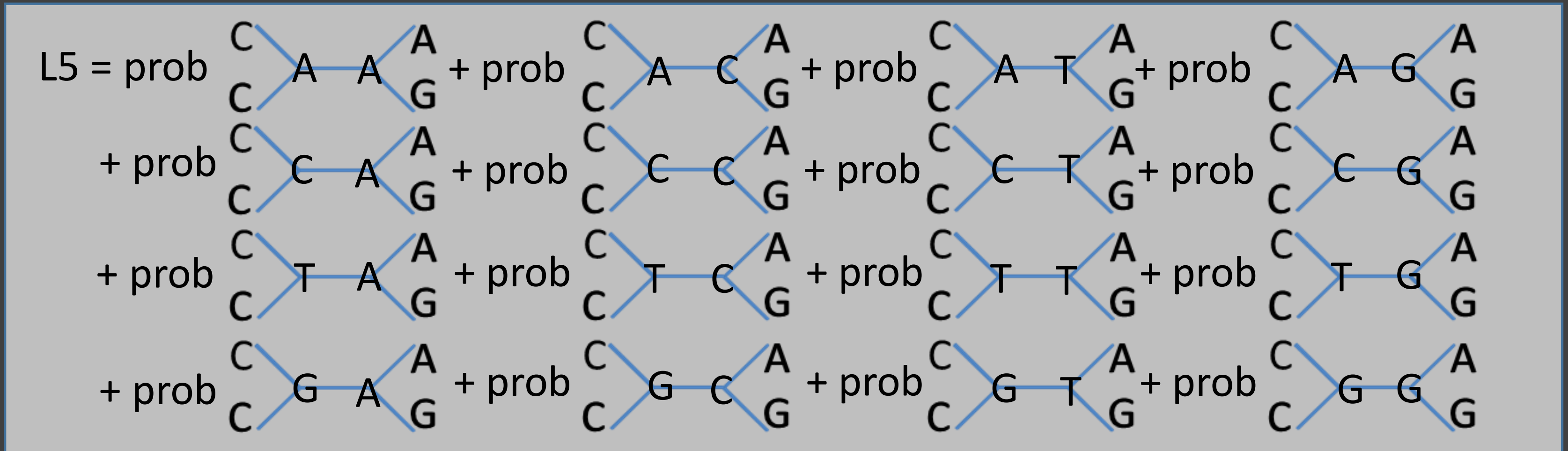
Maximum likelihood- based phylogenetic reconstruction from aligned sequence data



	1	2	3	4	5	6	7	8
OTU1	A	G	G	A	C	T	G	G
OTU2	A	A	C	T	C	T	G	C
OTU3	A	A	A	C	A	T	G	A
OTU4	A	A	A	G	G	T	G	A

See Figure 5.29 for application of ML to tree topology

1. starts with multiple sequence alignment
2. to compute L for given site (Ls), proceed over each site and sum probabilities of all possible sequence reconstructions for the each of the given sites for a given tree
3. to compute L for a given tree (Lt) take the product of all Ls for the given tree
4. or in practice take the sums of all ln(Ls)
5. search all tree topologies for maximum Lt



Total L for tree is sum of log L's for all sites (1-8)