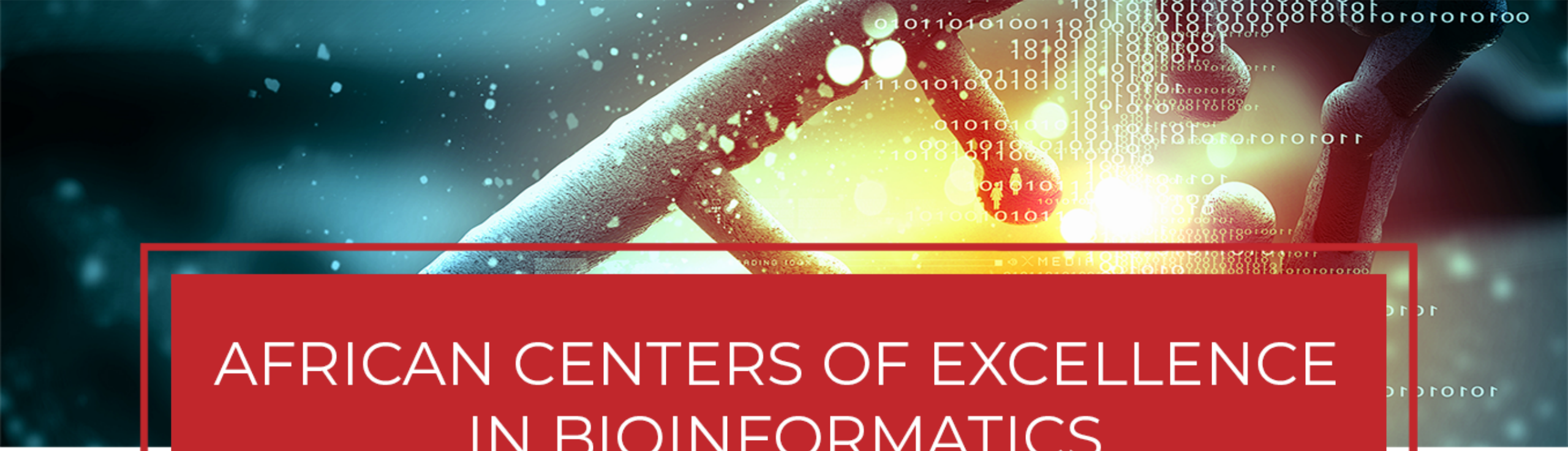
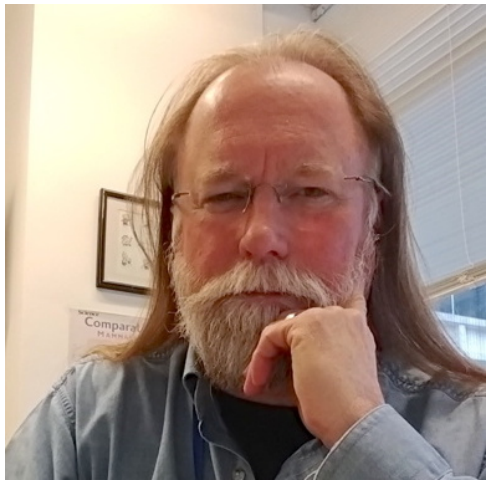


AFRICAN CENTERS OF EXCELLENCE IN BIOINFORMATICS

KAMPALA, UGANDA

CALCULATING EVOLUTIONARY DISTANCES AMONG SEQUENCES AND CORRECTION MODELS

Today's Instructor



Dr. Kurt Wollenberg, Ph.D. in Genetics

Ongoing Computational Biology projects:

- Hepatitis B molecular evolution
- CLAG protein family evolution

- Bioinformatics and Computational Biosciences Branch (BCBB), NIAID
- National Institutes of Health, Bethesda, MD USA.
- Contact our team via email:
  - Email: bioinformatics@niaid.nih.gov
  - Instructor: kurt.wollenberg@nih.gov

Class Materials

- Directory on Uganda ACE server:
  - File directory: user@kla-ac-bio-03:/home/bcbb_teaching_files
  - Large data files
- NIAID github repository:
  - https://github.com/niaid/ACE-2020
  - Code
  - Data files
  - Copies of lecture slides

GENETIC DIVERSITY

Nucleotide differences within a diploid population

	Sequences	# of pairwise differences							
Seq1	ATAAGGCTAGTCT	–							
Seq2	ATAAGGCAACTCT	2	–						
Seq3	ATAAGGCAACTCT	2	0	–					
Seq4	ATAAGGCTACTCT	1	1	1	–				
Seq5	ATACGGCTAGTCT	1	3	3	2	–			
Seq6	ATAAGGCTAGTCT	0	2	2	1	1	–		
Seq7	ATACGGCAACTCT	3	1	1	2	2	3	–	
Seq8	ATAAGGCTACTCT	1	1	1	0	2	1	2	–

Mean number of nucleotide differences

$$\Pi = \frac{1}{\left[\frac{n(n-1)}{2}\right]} \sum_{i<j} \Pi_{ij}$$

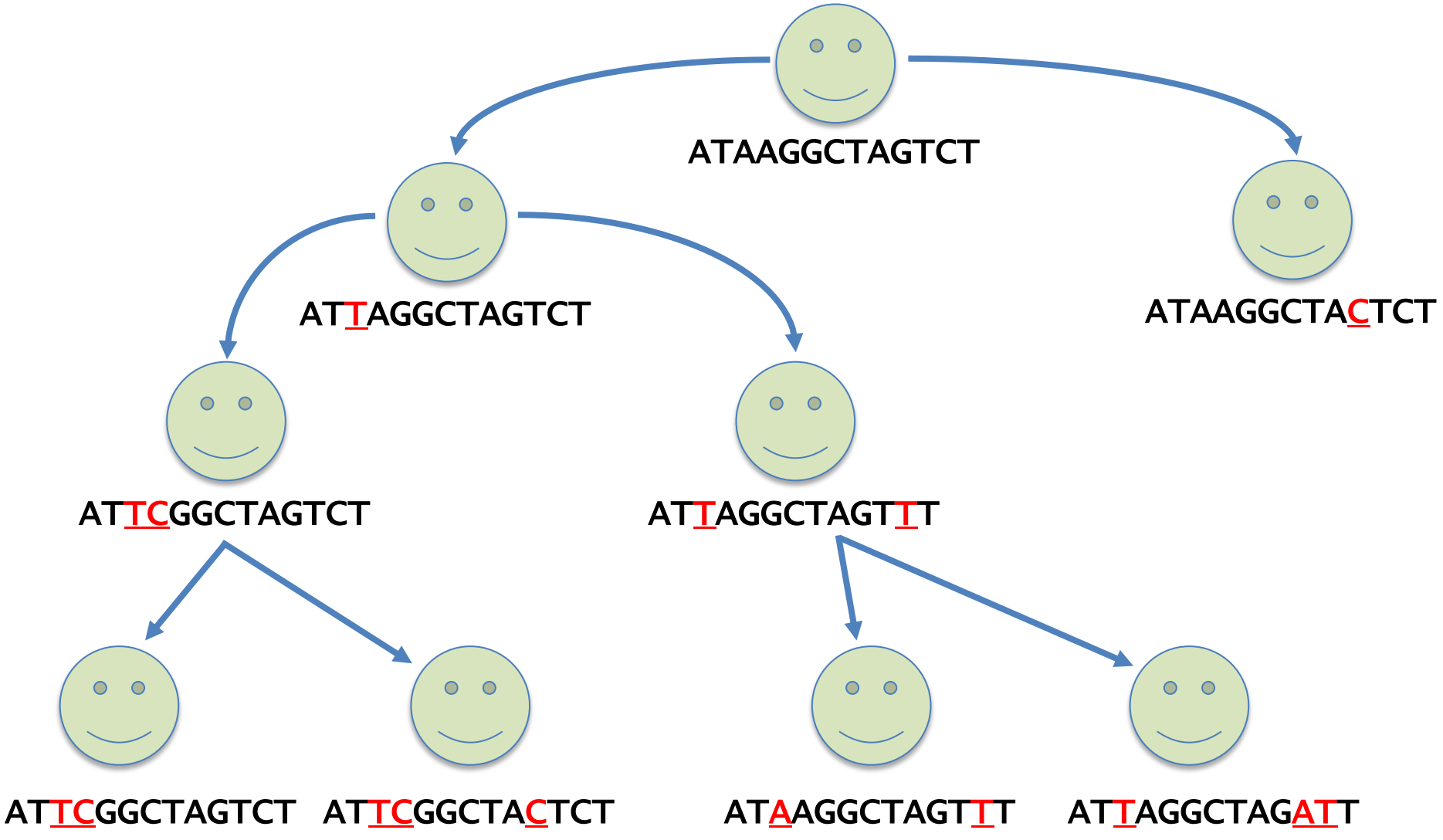
Π_{ij} is the number of nucleotide differences between sequences *i* and *j*.

Π = 42/28 = 1.5

EVOLUTIONARY DISTANCE

- How different (or similar) are our sequences?
- How do the evolutionary relationships of the sequences affect this?

EVOLUTIONARY DIVERSITY



GENETIC DIVERSITY

Nucleotide differences within a diploid population

	Sequences	# of pairwise differences								
Seq1	ATAAGGCTAGTCT	–								
Seq2	ATTAGGCTAGTCT	1	–							
Seq3	ATAAGGCTACTCT	1	2	–						
Seq4	ATTCGGCTAGTCT	2	1	3	–					
Seq5	ATTAGGCTAGTTT	2	1	3	2	–				
Seq6	ATTCGGCTAGTCT	2	1	3	0	2	–			
Seq7	ATTCGGCTACTCT	3	2	2	1	3	1	–		
Seq8	ATAAGGCTAGTTT	1	2	2	3	1	3	4	–	
Seq9	ATTAGGCTAGATT	3	2	4	3	1	3	4	2	–

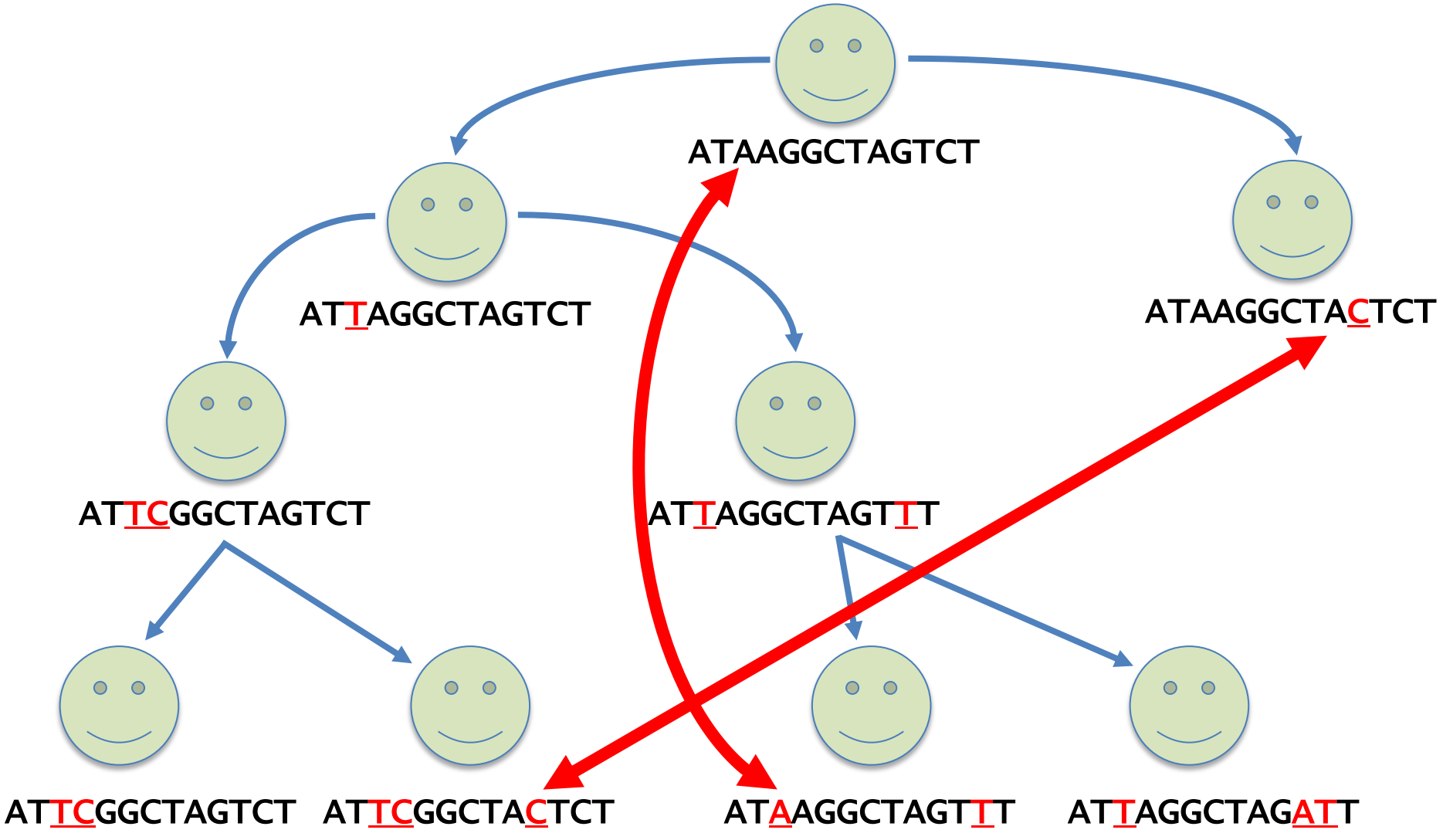
Mean number of nucleotide differences

$$\Pi = \frac{1}{\left[\frac{n(n-1)}{2}\right]} \sum_{i<j} \Pi_{ij}$$

Π_{ij} is the number of nucleotide differences between sequences *i* and *j*.

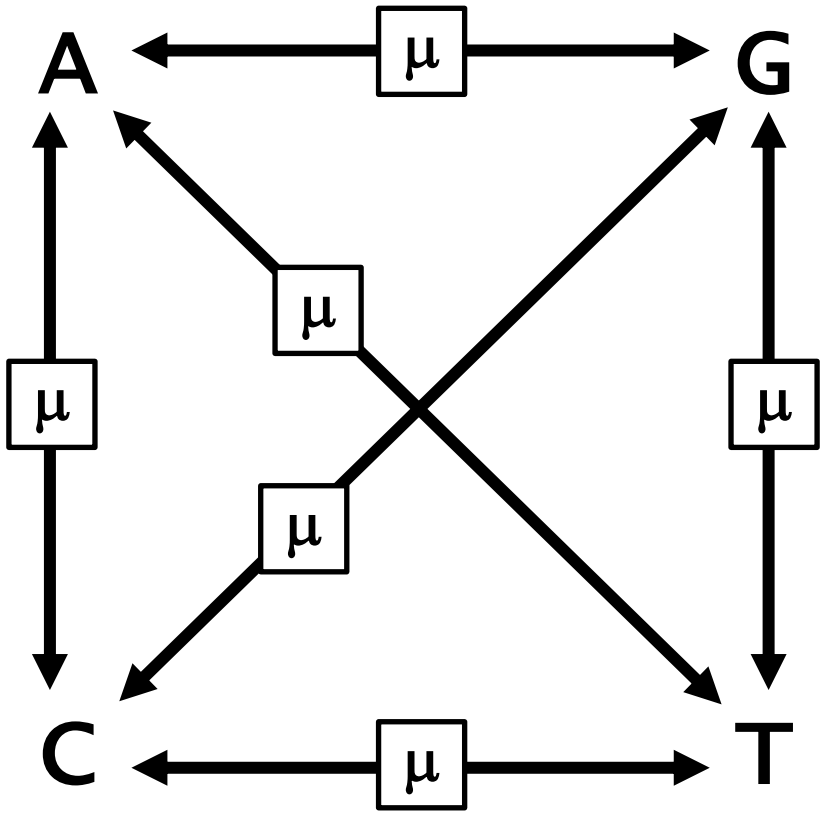
Π = 76/36 = 2.111

EVOLUTIONARY DIVERSITY



EVOLUTIONARY DIVERSITY

One-parameter substitution model (Jukes-Cantor [1969])



One uniform rate of substitution = μ

EVOLUTIONARY DIVERSITY

One-parameter substitution model (Jukes-Cantor [1969])

- At time t = 0, site has nucleotide A
- At t = 1, site has nucleotide A with probability $p_{A(1)} = 1 - 3\mu$
- At t = 2, site has nucleotide A with probability $p_{A(2)} = (1 - 3\mu)p_{A(1)} + \mu(1 - p_{A(1)})$
- In general, $p_{A(t+1)} = (1 - 3\mu)p_{A(t)} + \mu(1 - p_{A(t)})$
- $\Delta p_{A(t)} = p_{A(t+1)} - p_{A(t)} = -3\mu p_{A(t)} + \mu\left(1 - p_{A(t)}\right) = -4\mu p_{A(t)} + \mu$
- If change over time can be assumed to be continuous this is a first-order linear differential equation for $dp_{A(t)}/dt$ with the solution
- $p_{A(t)} = \frac{1}{4} + \left(p_{A(0)} - \frac{1}{4}\right)e^{-4\mu t} = \frac{1}{4} + \frac{3}{4}e^{-4\mu t}$, because $p_{A(0)} = 1$

EVOLUTIONARY DIVERSITY

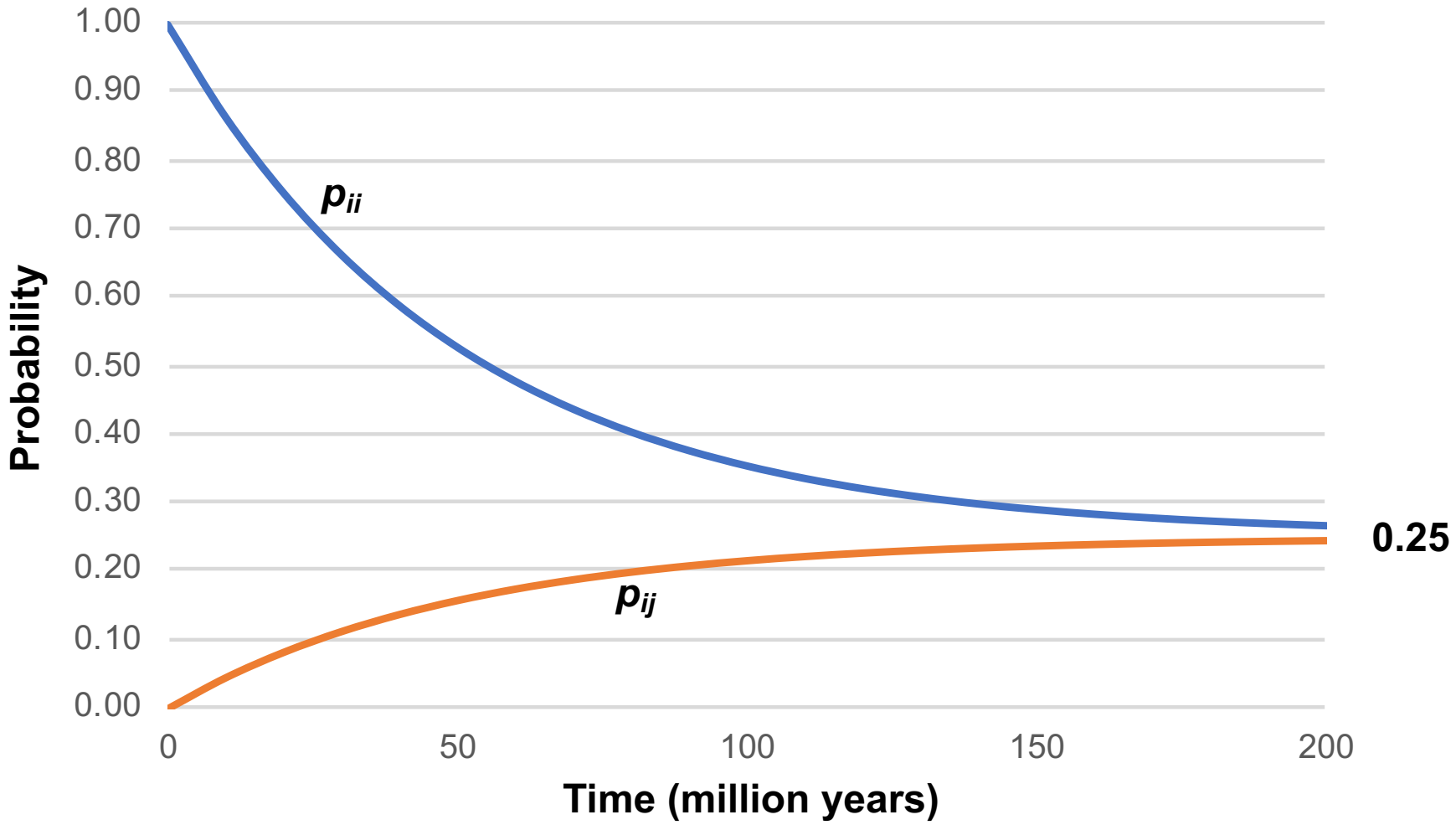
One-parameter substitution model (Jukes-Cantor [1969])

- $p_{A(t)} = \frac{1}{4} + \left(p_{A(0)} - \frac{1}{4}\right) e^{-4\mu t} = \frac{1}{4} + \frac{3}{4} e^{-4\mu t}$, because $p_{A(0)} = 1$
- If the initial nucleotide is not A, then $p_{A(0)} = 0$ and $p_{A(t)} = \frac{1}{4} - \frac{1}{4} e^{-4\mu t}$
- Under the assumptions of the Jukes-Cantor model, all nucleotides have the same probability of substitution, so $p_{AC} = p_{AG} = p_{AT} = p_{CG} = p_{CT} = p_{GT}$
- Giving the general substitution probabilities

$$p_{ii} = \frac{1}{4} + \frac{3}{4} e^{-4\mu t} \qquad p_{ij} = \frac{1}{4} - \frac{1}{4} e^{-4\mu t}$$

EVOLUTIONARY DIVERSITY

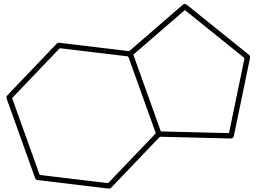
One-parameter substitution model (Jukes-Cantor [1969])

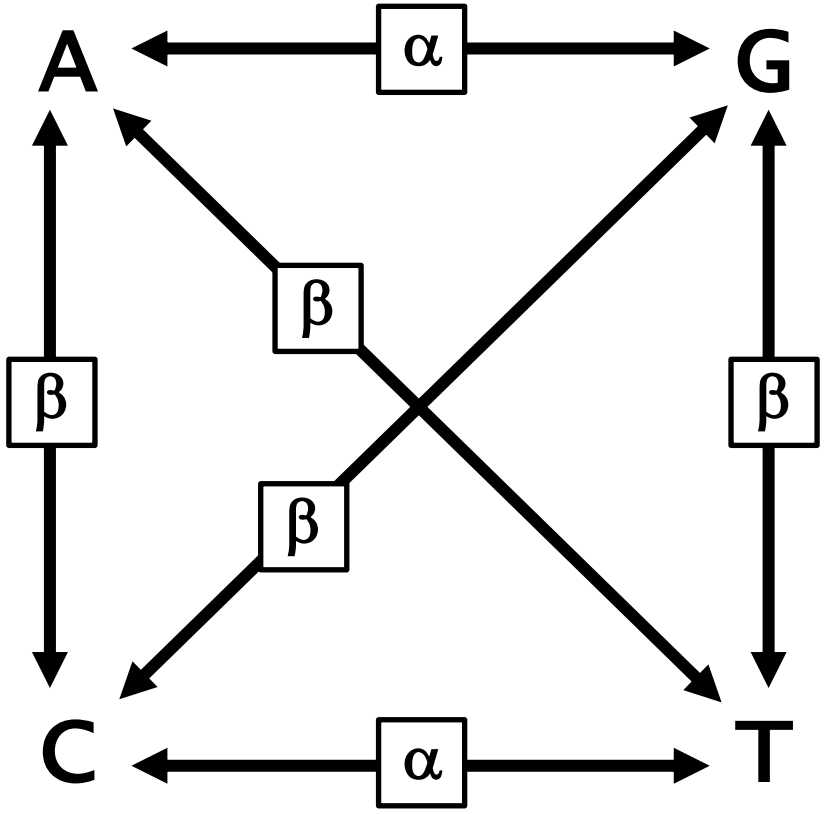


EVOLUTIONARY DIVERSITY

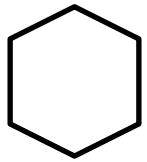
Two-parameter substitution model (Kimura [1980])

purines





pyrimidines



Two uniform rates of substitution:

transition = α

transversion = β

EVOLUTIONARY DIVERSITY

Two-parameter substitution model (Kimura [1980])

- At time *t* site has nucleotide A
- At *t* + 1, site has nucleotide A with probability

$$p_{A(t+1)} = (1 - \alpha - 2\beta)p_{A(t)} + \beta p_{C(t)} + \beta p_{T(t)} + \alpha p_{G(t)}$$

- Likewise, for a site having the other three nucleotides at time *t*

$$p_{C(t+1)} = \beta p_{A(t)} + (1 - \alpha - 2\beta)p_{C(t)} + \alpha p_{T(t)} + \beta p_{G(t)}$$

$$p_{T(t+1)} = \beta p_{A(t)} + \alpha p_{C(t)} + (1 - \alpha - 2\beta)p_{T(t)} + \beta p_{G(t)}$$

$$p_{G(t+1)} = \alpha p_{A(t)} + \beta p_{C(t)} + \beta p_{T(t)} + (1 - \alpha - 2\beta)p_{G(t)}$$

EVOLUTIONARY DIVERSITY

Two-parameter substitution model (Kimura [1980])

- For $p_{AA}(t) = p_{TT}(t) = p_{TT}(t) = p_{GG}(t)$: $p_{ii} = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t}$
- For transitional changes

$$p_{AG}(t) = p_{GA}(t) = p_{CT}(t) = p_{TC}(t): p_R = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t}$$

- For transversional changes

$$p_V = \frac{1}{4} - \frac{1}{4}e^{-4\beta t}$$

- Because there are two transversions, total probability is $p_{ii} + p_R + 2p_V = 1$

EVOLUTIONARY DIVERSITY

Two-parameter substitution model (Kimura [1980])

$$\mathbf{M} = \begin{pmatrix} 1-\alpha-2\beta & \beta & \beta & \alpha \\ \beta & 1-\alpha-2\beta & \alpha & \beta \\ \beta & \alpha & 1-\alpha-2\beta & \beta \\ \alpha & \beta & \beta & 1-\alpha-2\beta \end{pmatrix}$$

transitions

static

EVOLUTIONARY DIVERSITY

The general model

$$\mathbf{M} = \begin{pmatrix} 1-\alpha_{12}-\alpha_{13}-\alpha_{14} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & 1-\alpha_{21}-\alpha_{23}-\alpha_{24} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & 1-\alpha_{31}-\alpha_{32}-\alpha_{34} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & 1-\alpha_{41}-\alpha_{42}-\alpha_{43} \end{pmatrix}$$

EVOLUTIONARY DIVERSITY

Ancestral ATAAGGCTAGTCTTGCCATGG

Single Substitution	ATTAGGCTAGTCTTGCCATGG	ATAAGGCTACTCTTGCCATGG
Coincidental Substitution	ATTAGGCTTGTCTTGCCATGG	ATCAGGCTGCTCTTGCCATGG
	ATTGGGCTTGTCTTGCCATGG	ATCAGGCTGCTCTTGCCATGG
Multiple Substitution	ATTTGGCTTGTCTTGCCATGG	ATCAGGCTGCTCTTACCATGG
Parallel Substitution	ATTTGGCTTCTCTTGCCTTGG	ATCAGGCTGCTCTTACCTTGG
Back Substitution	ATTTGGCTTCTCATGCCTTGG	ATCAGGCTGCTCTTGCCTTGG
Convergent Substitution	ATTTGGCTTCTCGTGCCTTGG	ATCAGGCTGCTCGTGCCTTGG

EVOLUTIONARY DIVERSITY

ATTTGGCTTCTCGTGCCTTGG

ATCAGGCTGCTCGTGCCTTGG

Two sequences that are descended from a common ancestor *t* time units ago

How divergent are they?

Sequence Similarity – proportion identical

Sequence Diversity – proportion different

Diversity = 1 - Similarity

EVOLUTIONARY DIVERSITY

ATTTGGCTTCTCGTGCCTTGG

ATCAGGCTGCTCGTGCCTTGG

Two sequences that are descended from a common ancestor *t* time units ago

How identical are they?

- Consider a site has nucleotide *n* at time $t = 0$
- What is the probability at time $t > 0$ these two sequences have the same nucleotide?
- If no change, then this is p_{ii} for each sequence, so it is p_{ii}^2 for both
- If there have been multiple substitutions leading to the same nucleotide (parallel substitutions) there are three p_{ij} cases for each sequence, so three p_{ij}^2 probabilities

EVOLUTIONARY DIVERSITY

ATTTGGCTTCTCGTGCCTTGG

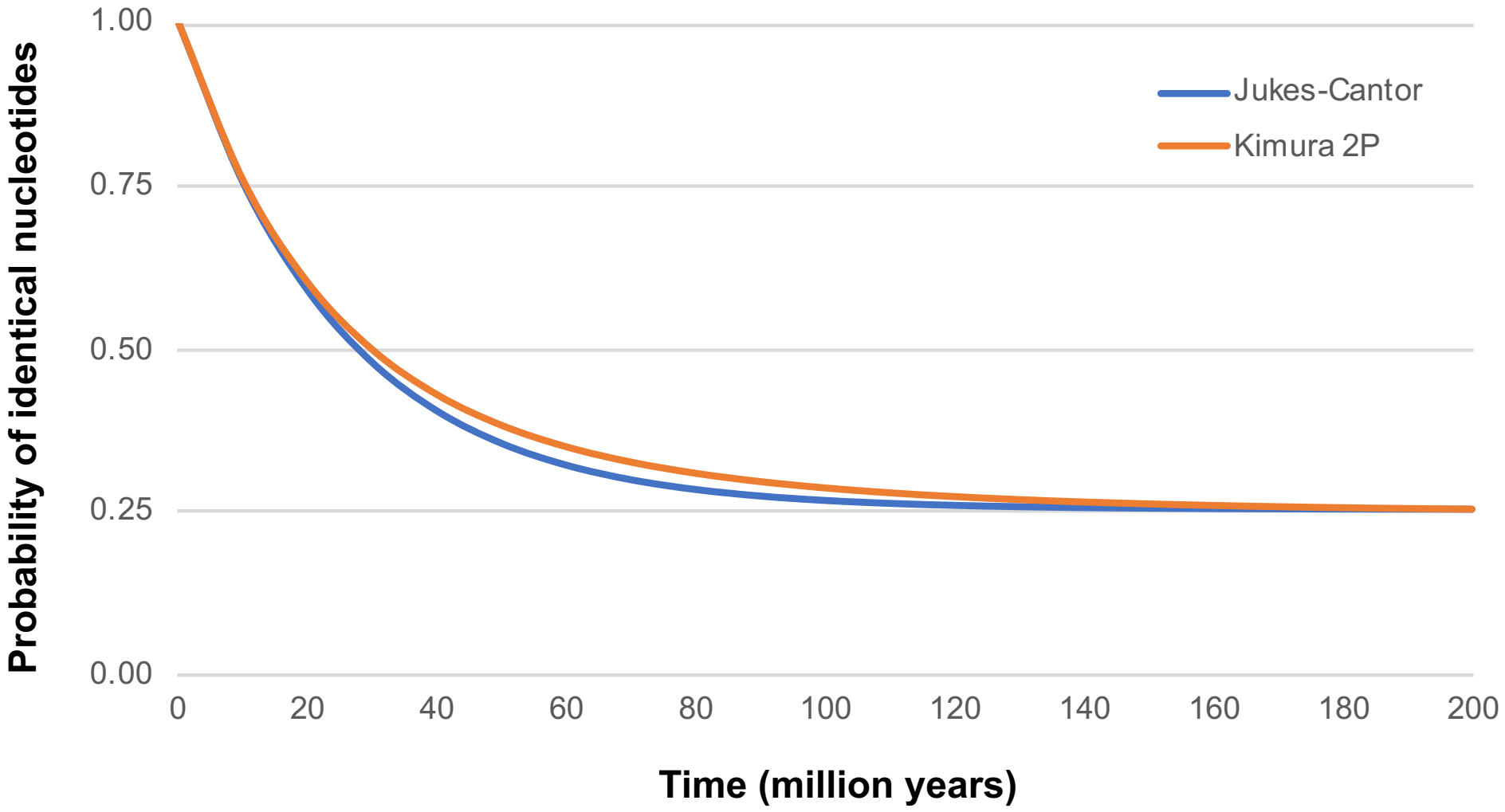
ATCAGGCTGCTCGTGCCTTGG

Two sequences that are descended from a common ancestor *t* time units ago

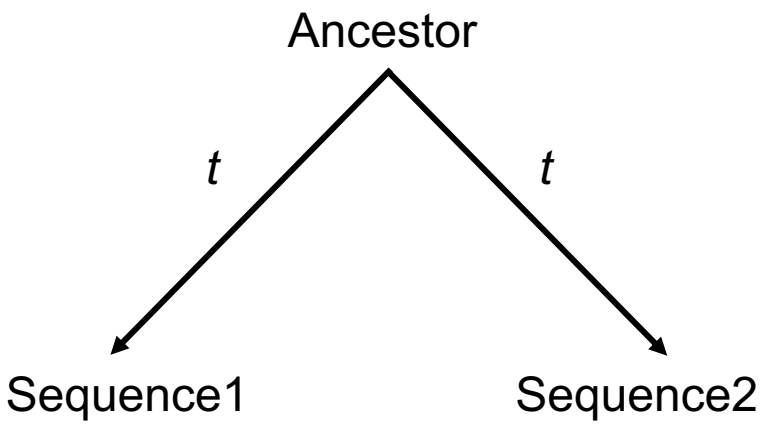
How identical are they?

- Therefore the probability of an identical nucleotide at this site in both sequences at time *t* is $I_{(t)} = p_{ii(t)}^2 + 3(p_{ij(t)}^2)$
- For the one-parameter substitution model we have $I_{(t)} = \frac{1}{4} + \frac{3}{4}e^{-8\mu t}$
- For the two-parameter substitution model we have $I_{(t)} = \frac{1}{4} + \frac{1}{4}e^{-8\beta t} + \frac{1}{2}e^{-4(\alpha+\beta)t}$
- Over long periods of time the expected identity between two sequences will reach an equilibrium (0.25 for these models) rather than 0.

EVOLUTIONARY DIVERSITY

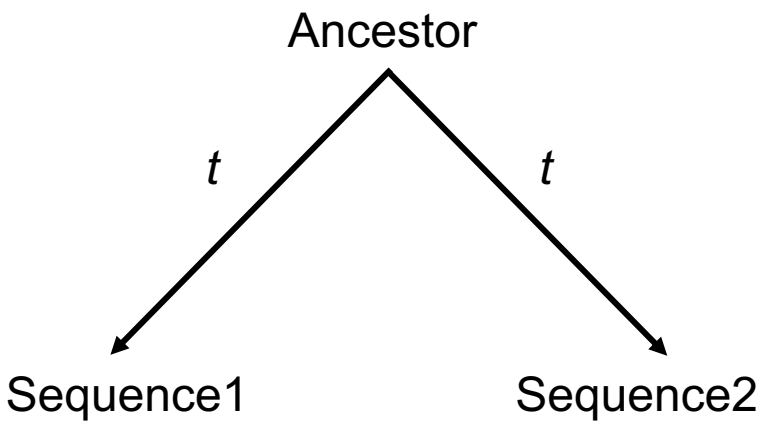


EVOLUTIONARY DISTANCE



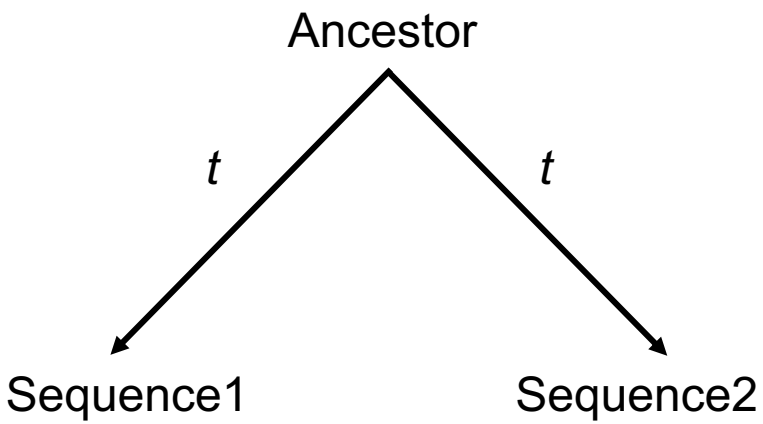
- Evolutionary distance expressed as number of substitutions per site *K*
- Rate of substitution varies based on the function of the DNA sequence
  - Coding vs noncoding (neutral)
  - If coding, synonymous variation vs nonsynonymous variation

EVOLUTIONARY DISTANCE



- For noncoding DNA sequence (same rate for all sites) probability two sequences are different at a site after time t is $p = 1 - I_{(t)}$ which is also the proportion of different sites between two sequences
- For the 1-parameter model $p = \frac{3}{4}(1 - e^{-8\mu t})$ so $8\mu t = -\ln(1 - \frac{4p}{3})$
- Since $K = 2(3\mu t)$ then $K = -(\frac{3}{4})\ln(1 - \frac{4p}{3})$ where K is the number of substitutions per site between the two sequences

EVOLUTIONARY DISTANCE

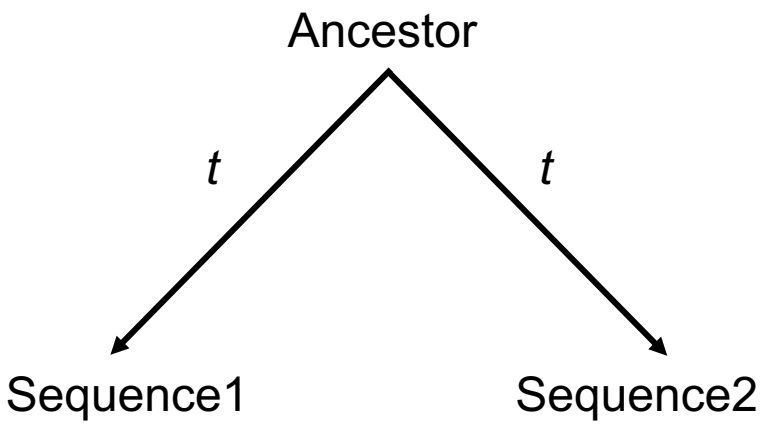


- For noncoding DNA sequence (same rate for all sites) probability two sequences are different at a site after time t is $p = 1 - I_{(t)}$
- For the 2-parameter model where *P* is the proportion of transitions and *Q* is the proportion of transversions

$$K = (\tfrac{1}{2})\ln(a) + (\tfrac{1}{4})\ln(b)$$

where $a = 1/(1-2P-Q)$ and $b = 1/(1-2Q)$

EVOLUTIONARY DISTANCE

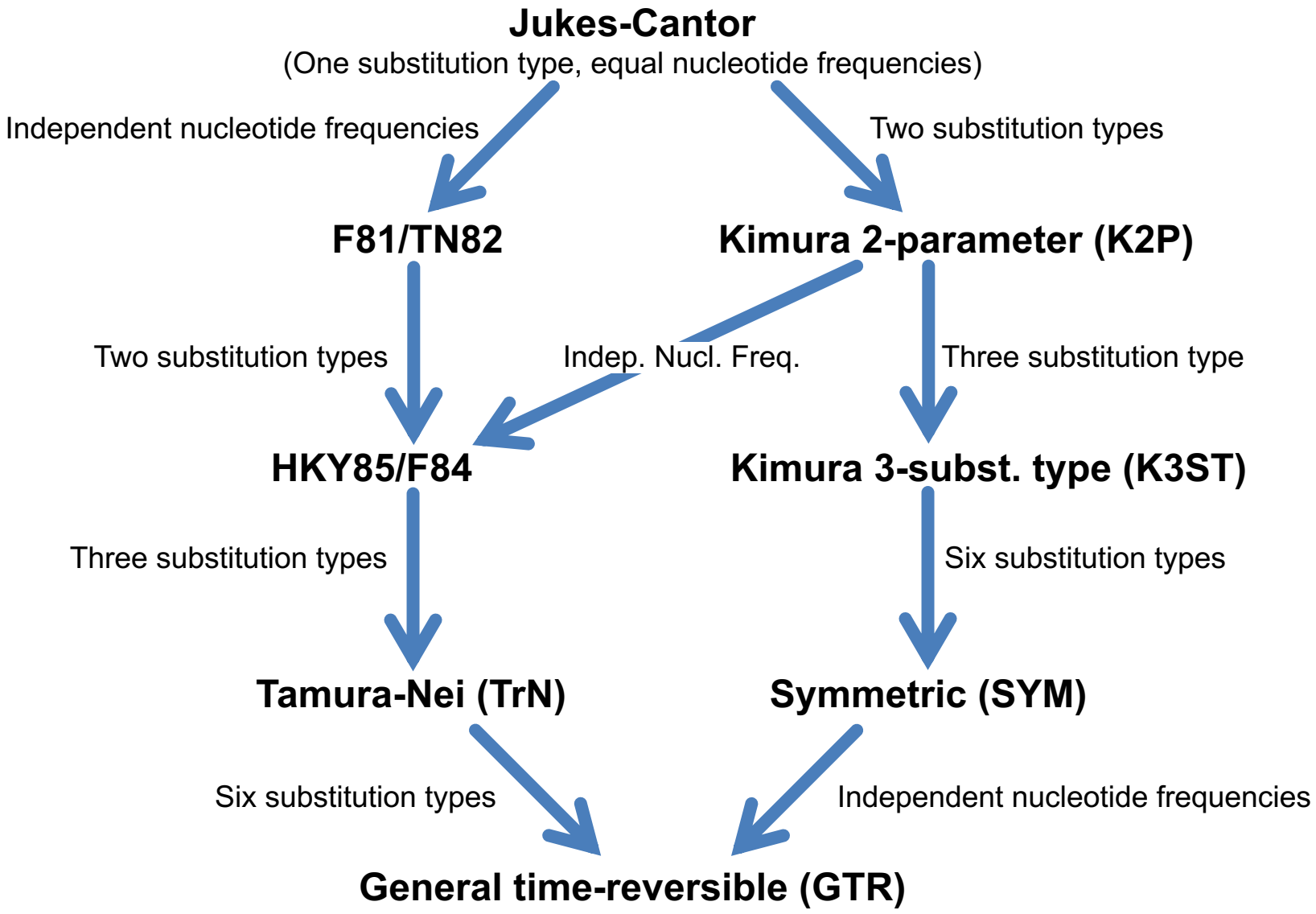


To relax the assumption of equal nucleotide frequencies in the sequences (the equilibrium frequencies) Tajima and Nei (1984) rewrote the Jukes-Cantor 1-parameter model as

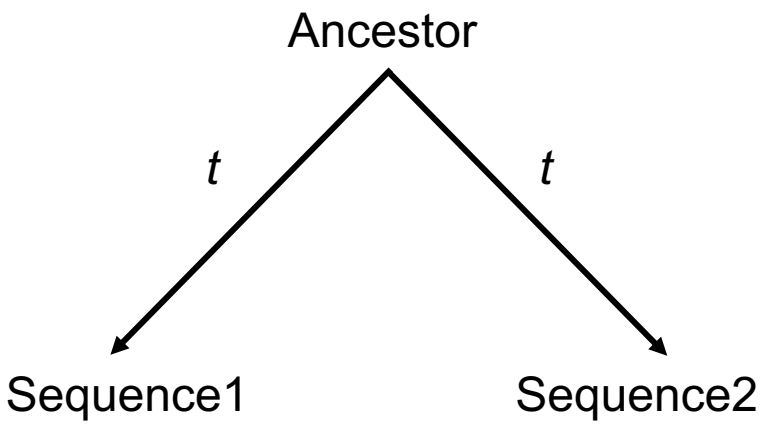
$$K = -b_1 \ln\left(1 - \frac{p}{b_1}\right)$$

where $b_1 = 1 - \sum q_i^2$ and the q_i are the equilibrium nucleotide frequencies, typically calculated from the data

SUBSTITUTION MODELS



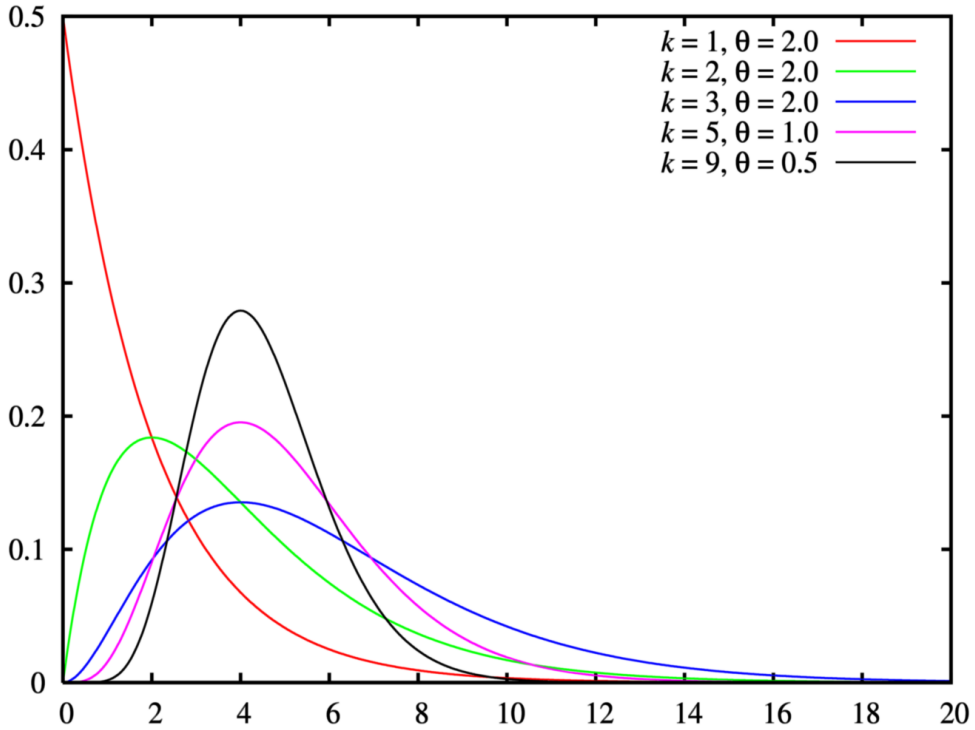
EVOLUTIONARY DISTANCE



- For a reasonable amount of diversity among sequences the assumption of uniform rates of substitution across sites is not valid
- Especially for coding sequences it is expected that there are different substitution rates at different sites
- Rates of synonymous and nonsynonymous substitution are expected to be different due to maintenance of protein structure and function

EVOLUTIONARY DISTANCE

- Substitution rates across sites are typically modeled with a discrete gamma distribution of four categories of substitution rates
- The continuous gamma distribution has a variable shape based on two parameters, shape and scale or rate (inverse scale).
- Different values of these parameters can make the distribution be more exponential or more normal.



EVOLUTIONARY DISTANCE

- OK, which model should I use to calculate my distances?
- Use software to evaluate how well your data fit different substitution models
  - MEGAX – Models -> Find best DNA/Protein model (ML)
  - jModeltest2 – Find the best fit DNA substitution model
  - protTest3 – Find the best fit protein sequence substitution model
- jModeltest2 and prottest3 available from https://github.com/ddarriba

EVOLUTIONARY DISTANCE

- MEGAX>Models and jModeltest2 use Bayesian Information Criteria (BIC) to evaluate model fit to data
- BIC is a function of likelihood of the model given the data, the number of parameters in the model, and the number of data points
- Lowest BIC = best fit model
- ΔBIC < 2: no difference in how these models fit the data
- ΔBIC >2, <6: good evidence that the best model is the best fit
- ΔBIC > 6, <10: strong evidence that the best model is the best fit
- ΔBIC > 10: I can’t even.

EVOLUTIONARY DISTANCE

Thank you for your time and attention