EE550
Computational Biology

Week 3 Course Notes

Instructor: Bilge Karaçalı, PhD

Topics

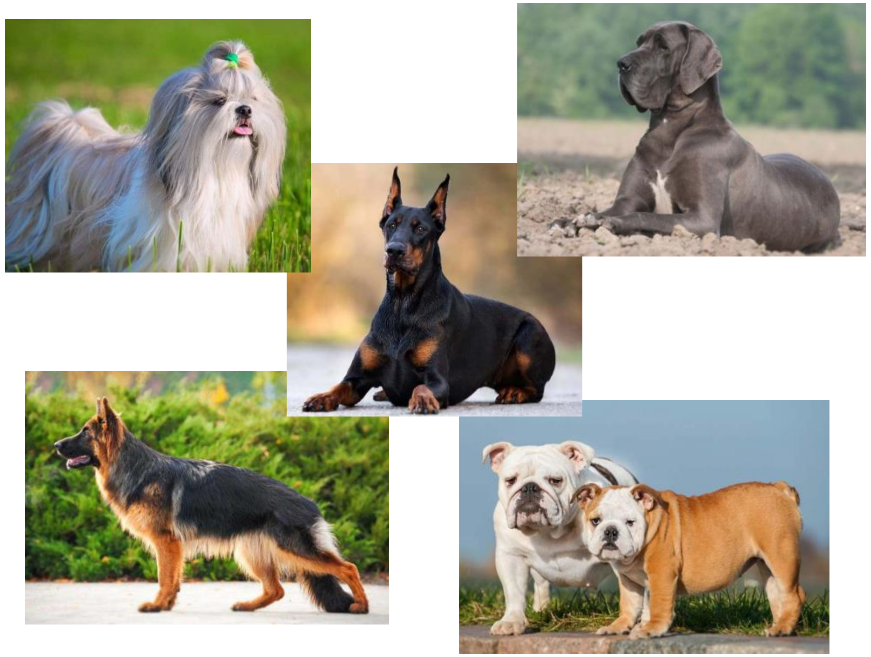
- Evolution mechanisms through mutations
  - Population genetics
  - Nucleic acid sequence evolution
    - Evolutionary distance vs. sequence distance
    - Jukes-Cantor model

Population Genetics

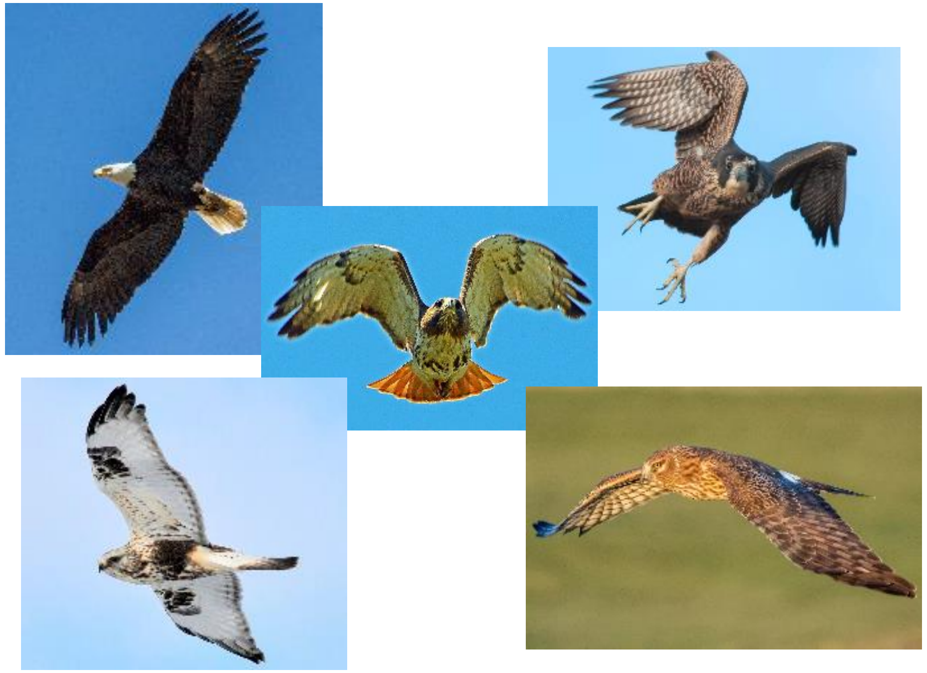
- Evolution:
  - Changes in the frequency as well as the sequence of genes in a population observed across time
  - Heritable changes in a population over many generations
  - …
- Two essential components:
  - Error-prone self replication produces genetic variants
  - Different variants incur varying levels of success at self replication through selection
    - Molecular evolution involves natural selection; selection carried out by nature
    - Unnatural selection; or artificial selection by humans forms the basis of agriculture
      - juicier and sweeter fruits
      - bigger and disease resistant crops
      - dogs and other animals bred selectively to fulfill different tasks

Case in Point: Dogs and Birds

- Dogs differ widely in their size and appearance, but belong to the same species
  - many years of selective breeding is responsible for all dog varieties
- Birds of prey look very similar but belong to different species



Source: http://www.dogbreedslist.info/all-dog-breeds/



Source: https://www.thespruce.com/types-of-birds-of-prey-387307

Nucleic Acid Sequences and Evolutionary History

- Organisms with common evolutionary ancestors share similar genetic sequences
  - At the time of genetic bifurcation, the two daughter species embark on different evolutionary paths
  - These different paths are characterized by the accumulation of different mutations
- The differences between their genetic sequences observed at the present time are related to the time of the bifurcation from the common ancestor
  - The earlier the separation, the higher the number of accumulated differences
  - The fraction of differences between sequences related to the evolutionary distances through mutation models

  → Estimation of the evolutionary relationship among a given set of genetic sequences from different organisms

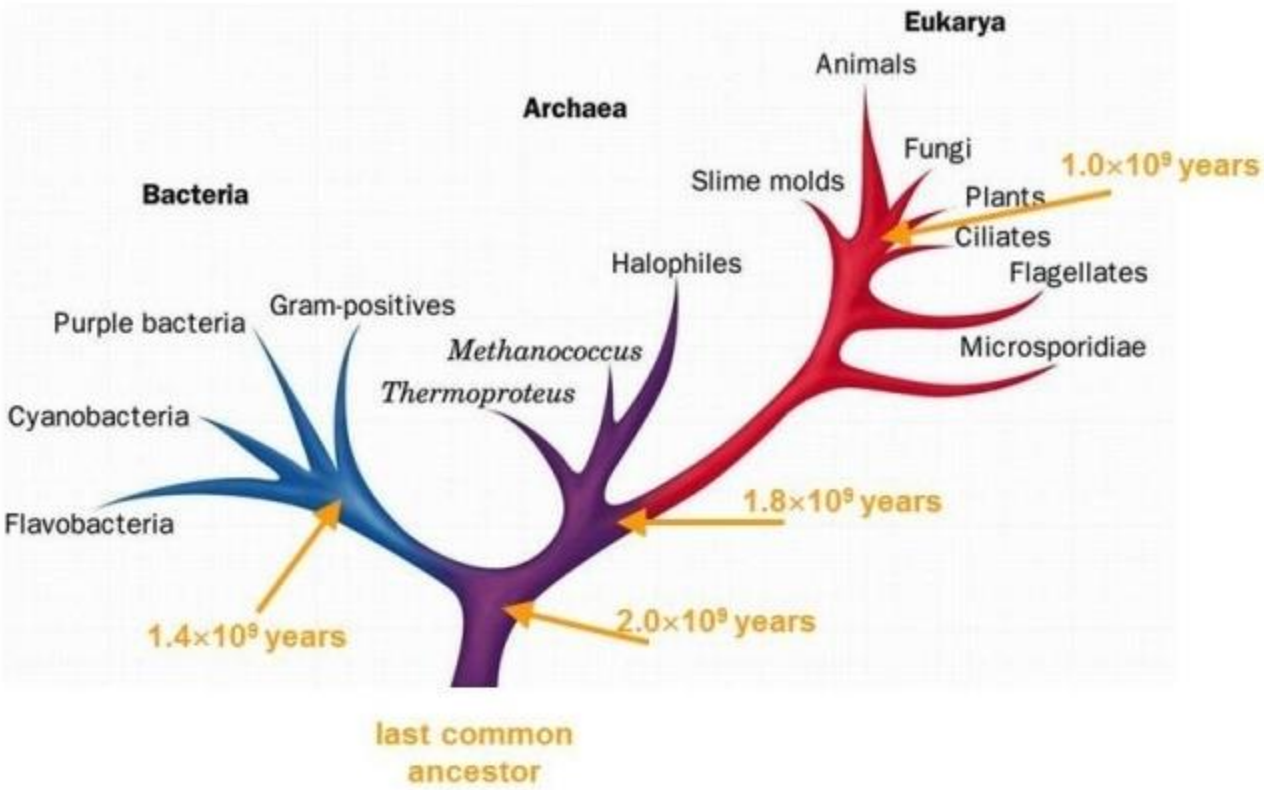
Spread of Mutations

- **An organism's fitness:** The ability to leave descendants in future generations
  - The greater the number of descendants, the higher the fitness
    - Has little to do with the health or the general well being of an organism
    - Has more to do with how beneficial its traits are in the organism's specific environment to leave descendants
- Mutations can have three types of effects on the fitness:
  - Advantageous: Increase the chance of leaving descendants
  - Neutral: No perceivable change in fitness
  - Deleterious: Decrease the chance of leaving descendants

Genetic Variation Between Species

- Evolution traces out **ancestors** and **descendants**
  - **Common ancestors of different species** from which they have diverged some time in the distant past
  - Some evolutionary tracts lead to survival
  - Other tracts disappear into extinction
    - Evolution is competition between alternative genetic configurations
    - Species that get outcompeted by others die out

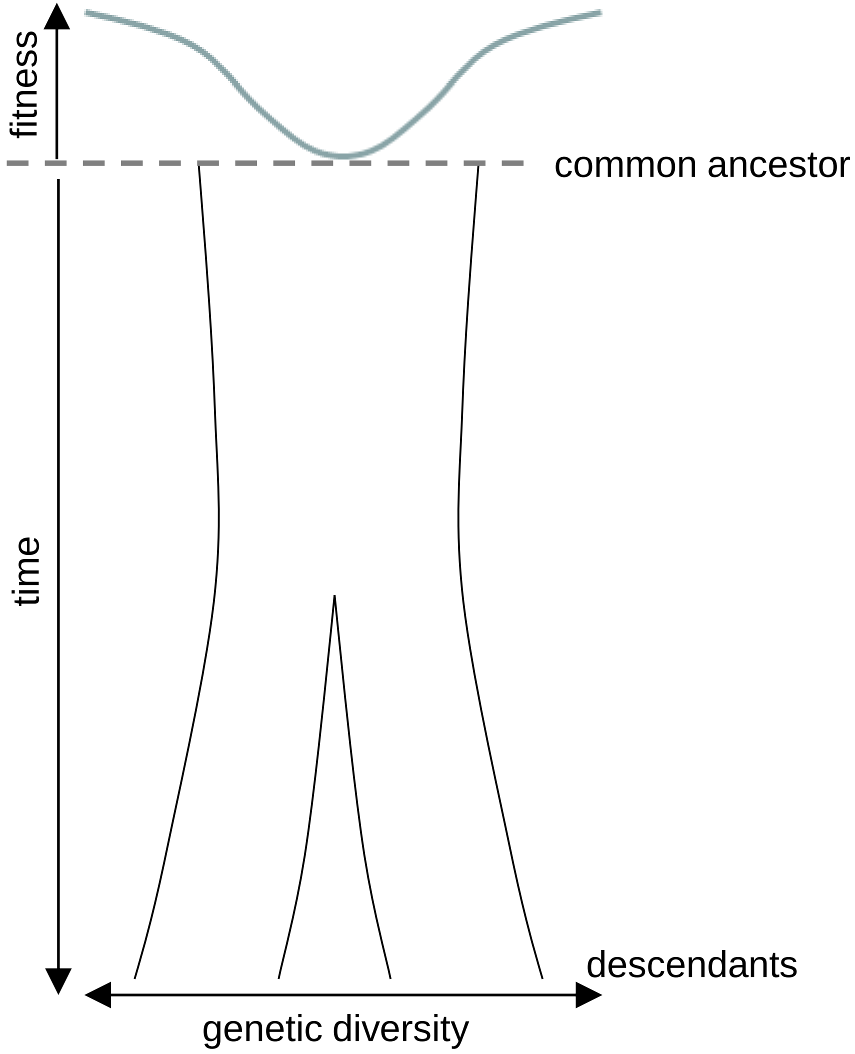
Tree of Life



Source: http://biologicalphysics.iop.org/cws/article/lectures/47042

Genetic Divergence Mechanism

- Changing environmental conditions work on the genetic variations within an ancestral species to create and shape the descendants
  - The descendants start off in the same species with slightly different genetic makeup
  - Time enhances the differences that allow exploiting different environmental niches
  - Eventually different species become “discernable”



Mutation Models on Nucleic Acid Sequences

- Genetic variations are characterized by differences in the **gene sequences**
  - Identical genes imply nearly identical organisms (up to chance effects from the environment)
  - Differences between organisms and species imply differences in their genes
- Quantification of these differences require

  stochastic models of nucleic acid sequence evolution
- These models also link **sequence differences** to evolutionary distances in units of **evolutionary time**
  - in terms of the nearest common ancestor in the evolutionary past

Modeling Nucleic Acid Substitutions

- **Objective**: to derive the relationship between the observed substitutions on different sequences and their evolutionary correspondence
  - Evolutionary correspondence refers to the amount of time in which the sequences went down independent evolutionary paths
- **Premises**:
  - Substitutions occur randomly
  - Fixation is assumed to have been…
    - achieved when comparing sequences of different species
    - not achieved when comparing sequences across individuals
  - Rates of substitution are constant for the sequences involved during the corresponding time period
- **Approach**:
  - Establish a functional relationship between a sequence distance and the corresponding evolutionary distance

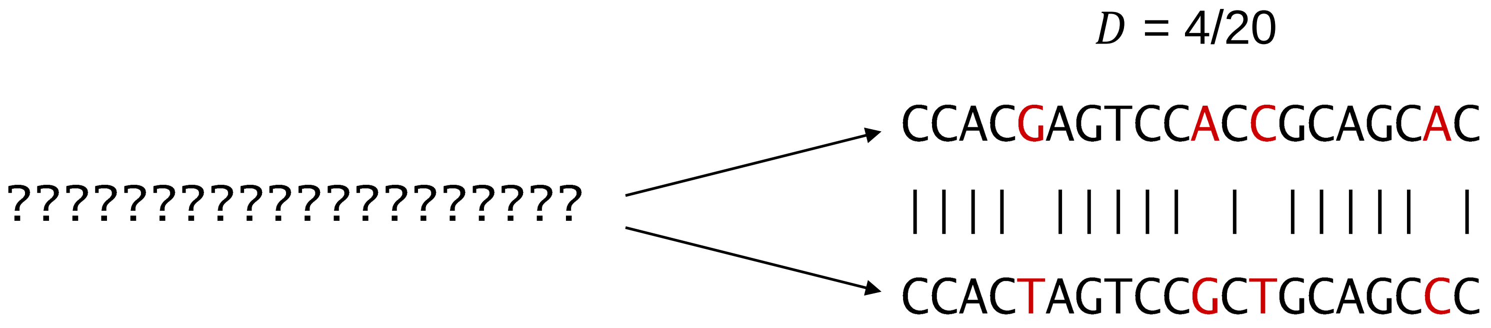
$$\text{evolutionary distance (in time units)} = \mathcal{F}(\text{sequence distance})$$

Modeling Nucleic Acid Substitutions

- Sequence difference D:
  - Measured by the **fraction of nucleotides that are different** between two nucleic acid sequence fragments
  - Correlates linearly with the evolutionary time span for small time periods, but varies nonlinearly for large time periods
  - Can be measured quantitatively for any given two sequences simply by counting the number of sites where the sequences do not match
    - Hamming distance in coding theory
- Evolutionary distance d:
  - Measured by the **average number of substitutions** that have occurred per site between the two sequences during the time span of independent evolution
  - Correlates linearly with the time span of independent evolution for all time ranges, small **AND** large
  - Cannot be measured directly but can be inferred from D using a stochastic model

Modeling Nucleic Acid Substitutions

- Visible substitutions:
  - One sequence remains the same and the other incurs a substitution, or
  - Both sequences incur substitutions into different nucleotides
- Invisible substitutions:
  - Neither sequence incurs a substitution (i.e., the original nucleotide remains preserved/conserved in both sequences), or
  - Both sequences incur substitutions into the same nucleotide
- Annulled substitutions:
  - Successive substitutions in both sequences result in the same nucleotide



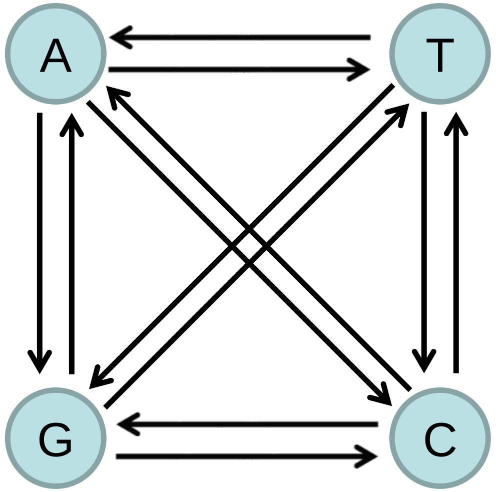
D = 4/20

```
                       CCACGAGTCCACCGCAGCAC
????????????????????   |||| ||||| | ||||| |
                       CCACTAGTCCGCTGCAGCCC
```

The Jukes-Cantor Model

- The substitution phenomenon is modeled by a Markov chain
- In the Jukes-Cantor model, the rate of substitution from one base to any other is denoted by α, in number of substitutions per unit time
  - Thus, the net rate of change of a base is 3α
  - $\alpha \ll 1$
- The corresponding state transition rate matrix is given by

$$Q = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}$$



The Jukes-Cantor Model

- The resulting transition probability matrix is

$$P(t) = e^{Qt} = \begin{bmatrix} \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} \\ \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} \\ \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} \\ \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} \end{bmatrix}$$

or, more simply,

$$P_{i,j}(t) = \begin{cases} \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} & \text{if } i = j \\ \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \text{otherwise} \end{cases}$$

The Jukes-Cantor Model

- Note that $P_{i,j}(t)$ represents the probability with which the i’th nucleotide occupying a specific site on the original DNA sequence will be replaced by the j’th nucleotide in t units of time
- This allows calculating the average sequence difference between the original sequence and the evolving sequence as the expected value

$$D(t) = \sum_{i,j} \mathbf{1}(i \neq j) P_{i,j}(t)\pi_i$$

- Assuming an equal rate of nucleotides across the DNA, i.e. $\pi_i = 1/4$ for all $i = 1,2,3,4$, we get

$$D(t) = 12\left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)\frac{1}{4} = \frac{3}{4} - \frac{3}{4}e^{-4\alpha t}$$

- In addition, the incurred evolutionary distance by the evolving sequence to the original sequence is given by

$$d(t) = \sum_{i,j} \mathbf{1}(i \neq j) Q_{i,j} t \pi_i = 3\alpha t$$

The Jukes-Cantor Model

- To relate the observed sequence distance D between two evolved sequences to the evolutionary distance d between them:
  - the first sequence incurs $3\alpha t$ from the original
  - the second sequence incurs another $3\alpha t$ from the original, independent of the substitutions of the first one
  - this implies a total evolutionary distance of

$$d(t) = 6\alpha t$$

between the independently evolving sequences

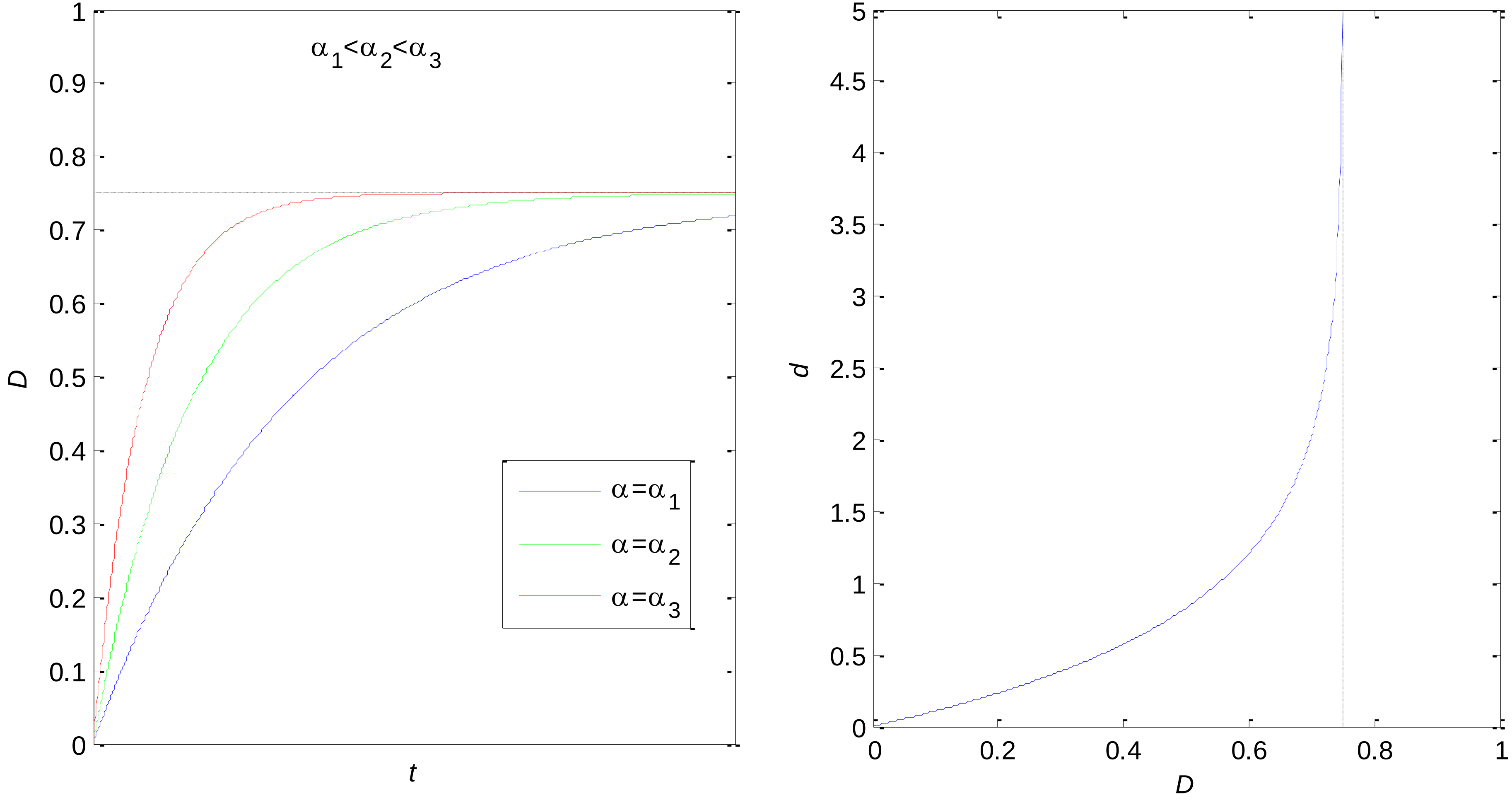
  - furthermore, a combined evolution time of $2t$ produces a sequence distance of

$$D(t) = \frac{3}{4} - \frac{3}{4} e^{-8\alpha t}$$

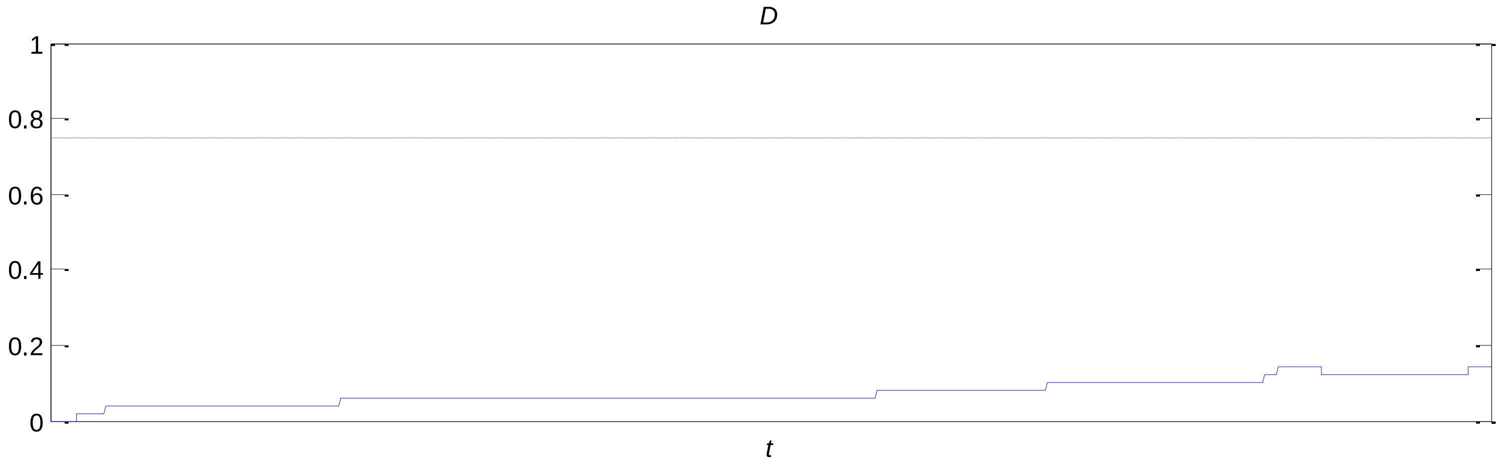
  - solving for the two in terms of αt, we get

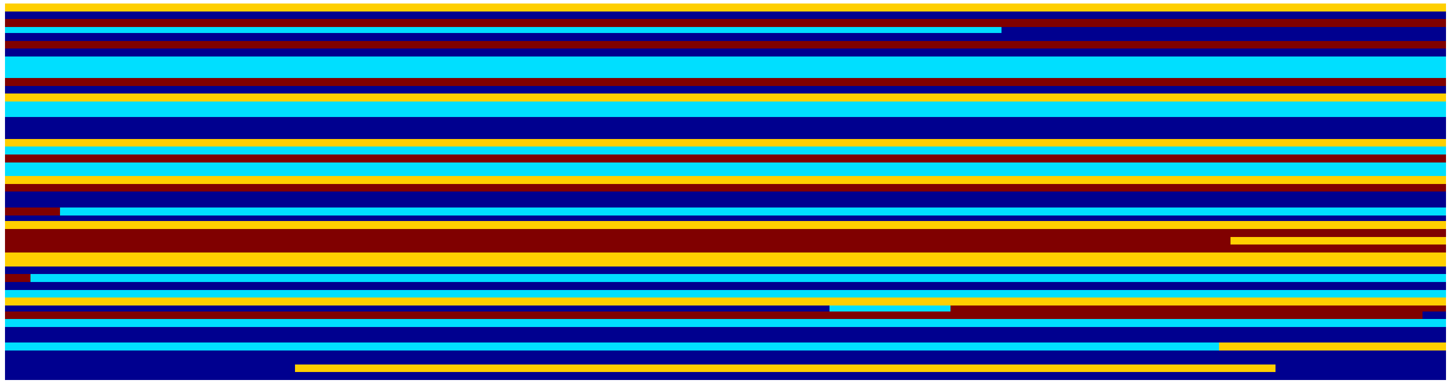
$$D = \frac{3}{4} - \frac{3}{4} e^{-\frac{4}{3}d} \quad \text{or} \quad d = -\frac{3}{4} \log\left(1 - \frac{4}{3} D\right)$$

The Jukes-Cantor Model

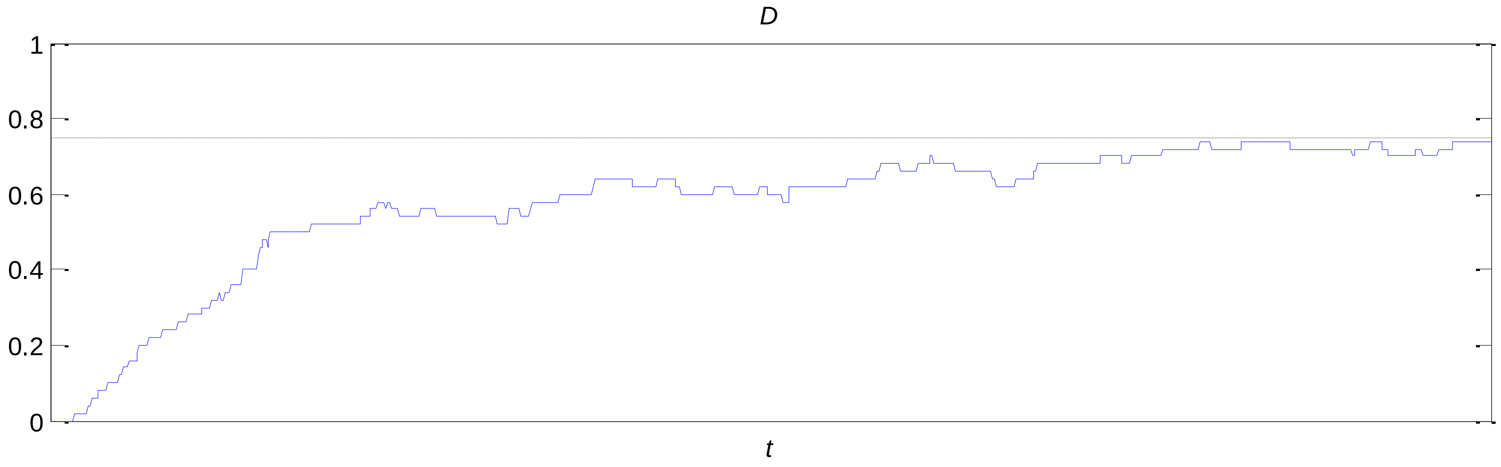


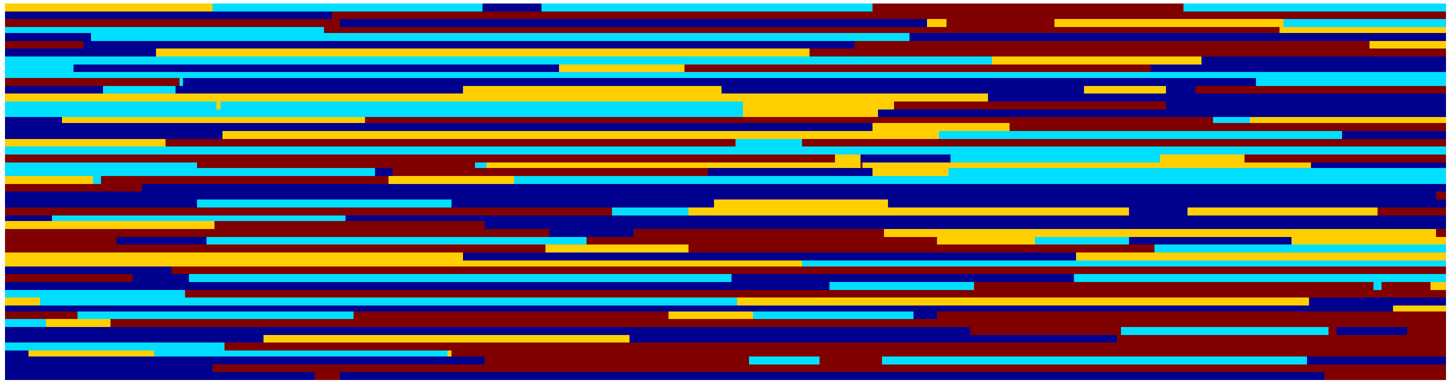
Example: Slow Evolution of a Single Sequence



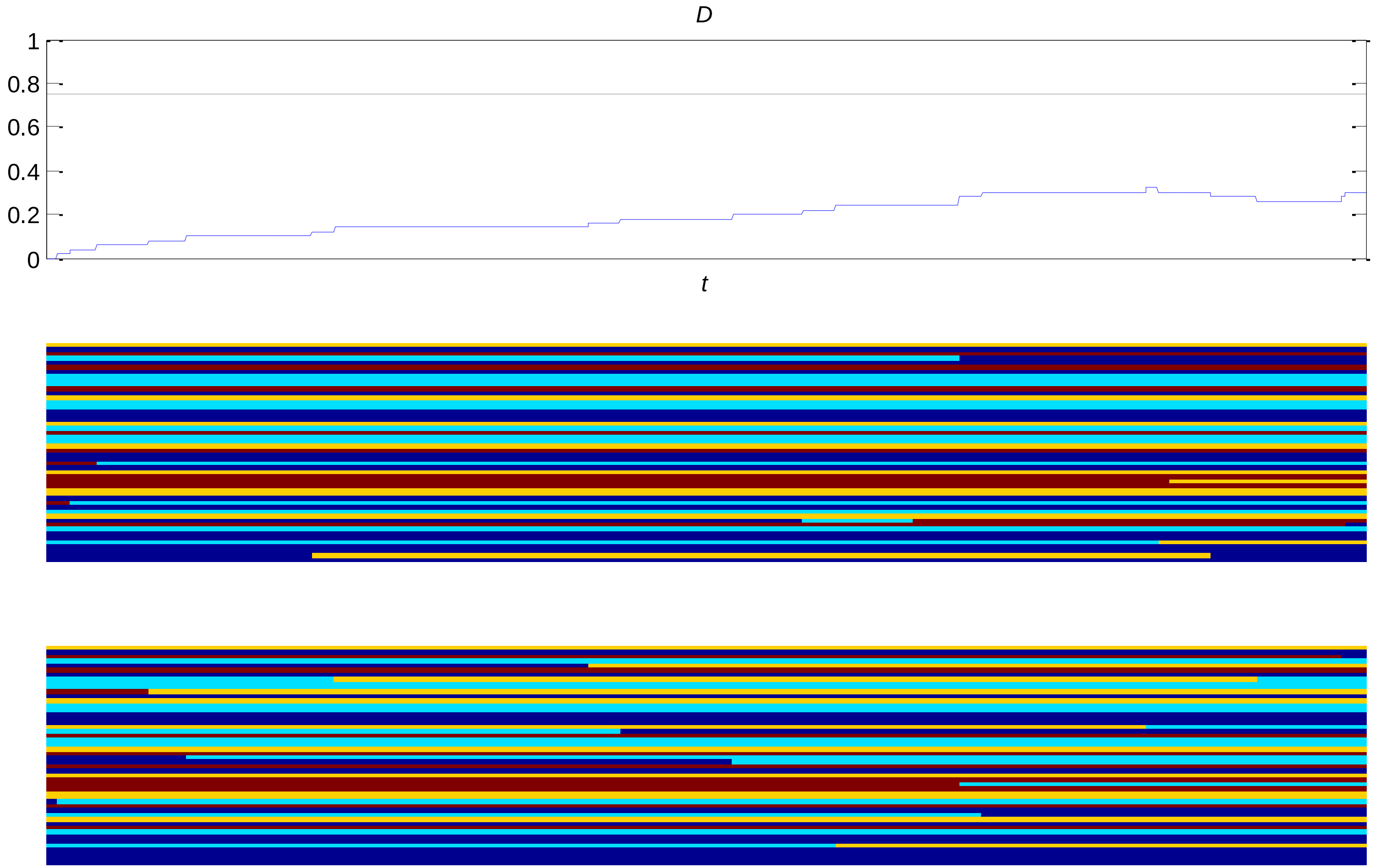


Example: Fast Evolution of a Single Sequence





Example: Simultaneous Evolution of Two Sequences



Alternative Models

Jukes-Cantor				
	A	G	C	T
A	$\star$	α	α	α
G	α	$\star$	α	α
C	α	α	$\star$	α
T	α	α	α	$\star$

HKY				
	A	G	C	T
A	$\star$	$\alpha\pi_G$	$\beta\pi_C$	$\beta\pi_T$
G	$\alpha\pi_A$	$\star$	$\beta\pi_C$	$\beta\pi_T$
C	$\beta\pi_A$	$\beta\pi_G$	$\star$	$\alpha\pi_T$
T	$\beta\pi_A$	$\beta\pi_G$	$\alpha\pi_C$	$\star$

Kimura 2-parameter				
	A	G	C	T
A	$\star$	α	β	β
G	α	$\star$	β	β
C	β	β	$\star$	α
T	β	β	α	$\star$

General Reversible				
	A	G	C	T
A	$\star$	$\alpha_{A\to G}$	$\alpha_{A\to C}$	$\alpha_{A\to T}$
G	$\alpha_{G\to A}$	$\star$	$\alpha_{G\to C}$	$\alpha_{G\to T}$
C	$\alpha_{C\to A}$	$\alpha_{C\to G}$	$\star$	$\alpha_{C\to T}$
T	$\alpha_{T\to A}$	$\alpha_{T\to G}$	$\alpha_{T\to C}$	$\star$

Variable Substitution Rates

- The Jukes-Cantor model as well as the more sophisticated ones assume that all sites along the DNA are equally prone to base substitutions
  - $P_{i,j}(t)$ is assumed to be the same regardless of the position of the nucleotide on the sequence
- This assumption simplifies the analysis, but does not exactly hold in reality
  - Some sites are structurally or functionally important, and evolve more slowly
    - Due to strong selective pressure
    - Some very important sites are practically invariant

Variable Substitution Rates

- Relaxing this assumption requires incorporating site-specific variation in observed differences
  - Jukes-Cantor model with a fixed fraction q of invariable sites:

$$d = -\frac{3}{4}(1-q)\log\left(1 - \frac{4D}{3-3q}\right)$$

  - Jukes-Cantor model where the variability of sites is governed by a gamma distribution:

$$d = \frac{3}{4}a\left(\left(1 - \frac{4}{3}D\right)^{-1/a} - 1\right)$$

where a is the shape parameter of the gamma distribution governing the probability of a site being subject to a substitution rate of r, described by the probability density function

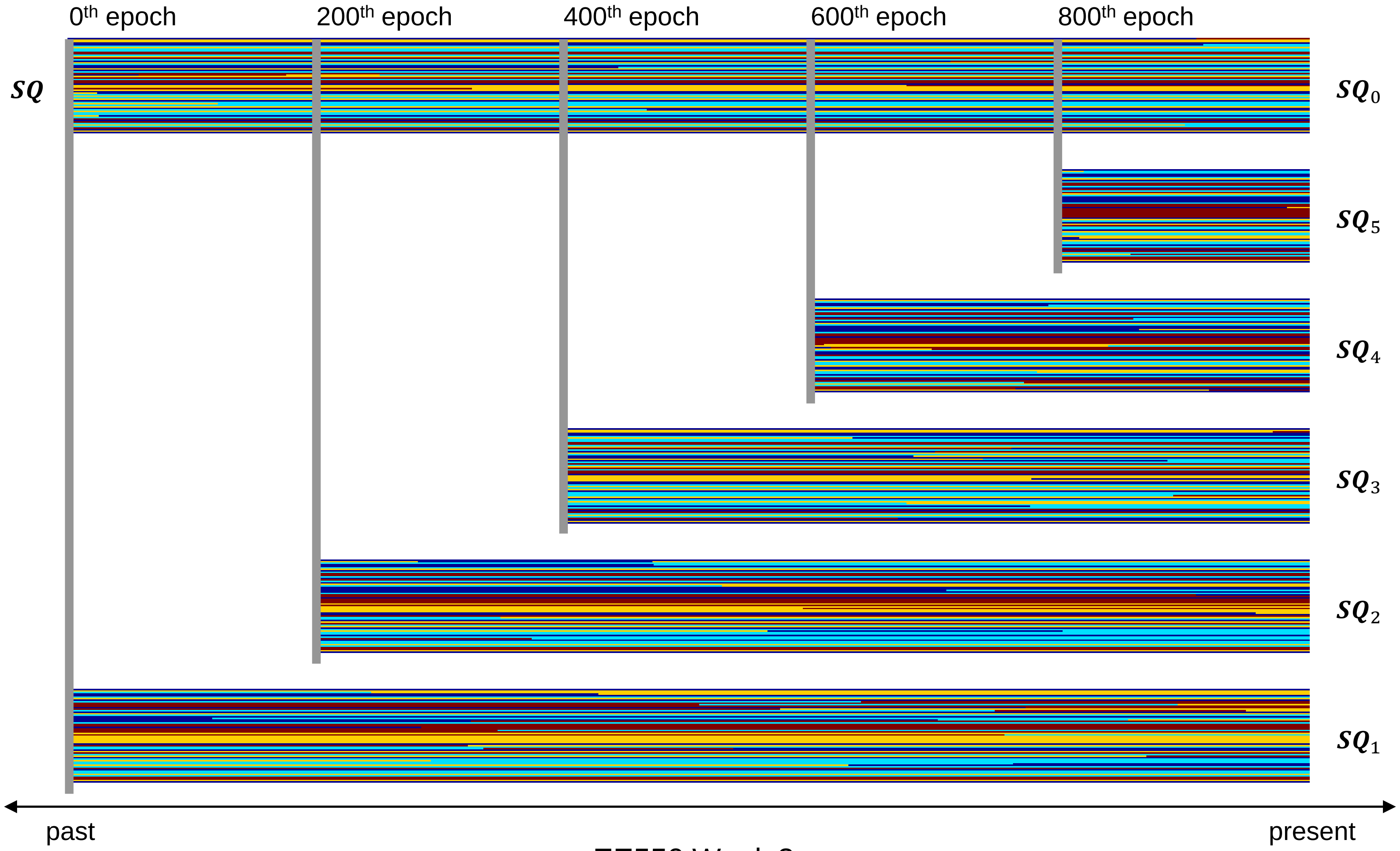
$$f_R(r;a) = Zra^{-1}e^{-ar}$$

Example: Evolutionary Siblinghood

- Data
  - A random “original” nucleic acid sequence $\boldsymbol{SQ}$ of length $N = 100$ nucleotides undergoing point mutations according to a Jukes-Cantor model
  - Molecular evolution carried out *in silica* for 1000 epochs
    - $\boldsymbol{SQ}^{(k)}$: The evolved sequence at the k’th epoch
    - $\boldsymbol{SQ}^{(0)} = \boldsymbol{SQ}$ (the original sequence)
    - $\boldsymbol{SQ}_0 = \boldsymbol{SQ}^{(1000)}$
  - A total of 5 sibling sequences, $\boldsymbol{SQ}_1$, $\boldsymbol{SQ}_2$, $\boldsymbol{SQ}_3$, $\boldsymbol{SQ}_4$ and $\boldsymbol{SQ}_5$ identified as
    - $\boldsymbol{SQ}_1^{(0)} = \boldsymbol{SQ}^{(0)}$, $\boldsymbol{SQ}_1 = \boldsymbol{SQ}_1^{(1000)}$
    - $\boldsymbol{SQ}_2^{(0)} = \boldsymbol{SQ}^{(200)}$, $\boldsymbol{SQ}_2 = \boldsymbol{SQ}_2^{(800)}$
    - $\boldsymbol{SQ}_3^{(0)} = \boldsymbol{SQ}^{(400)}$, $\boldsymbol{SQ}_3 = \boldsymbol{SQ}_3^{(600)}$
    - $\boldsymbol{SQ}_4^{(0)} = \boldsymbol{SQ}^{(600)}$, $\boldsymbol{SQ}_4 = \boldsymbol{SQ}_4^{(400)}$
    - $\boldsymbol{SQ}_5^{(0)} = \boldsymbol{SQ}^{(800)}$, $\boldsymbol{SQ}_5 = \boldsymbol{SQ}_5^{(200)}$

  evolved independently through the remaining epochs.
- Procedure:
  - Compute the sequence distances $D_{0,j}$ between $\boldsymbol{SQ}_0$ and $\boldsymbol{SQ}_1$, $\boldsymbol{SQ}_2$, $\boldsymbol{SQ}_3$, $\boldsymbol{SQ}_4$ and $\boldsymbol{SQ}_5$
  - Calculate the evolutionary distances $d_{0,j}$ from $D_{0,j}$ using the Jukes-Cantor model

Example: Sequence Data



Example: Sequence Data

$\boldsymbol{SQ_0}$

AGTACCCGGGGCCATCGAAG...

$\boldsymbol{SQ_1}$

ATTTCCCGTCGAGATCGAAT...

$\boldsymbol{SQ_2}$

ATTACCCGTTGCGAGGGAAG...

$\boldsymbol{SQ}$

ATTACCCGTGGCGATCGATG...

$\boldsymbol{SQ_3}$

AGTACACGTGGCAATCGAGG...

$\boldsymbol{SQ_4}$

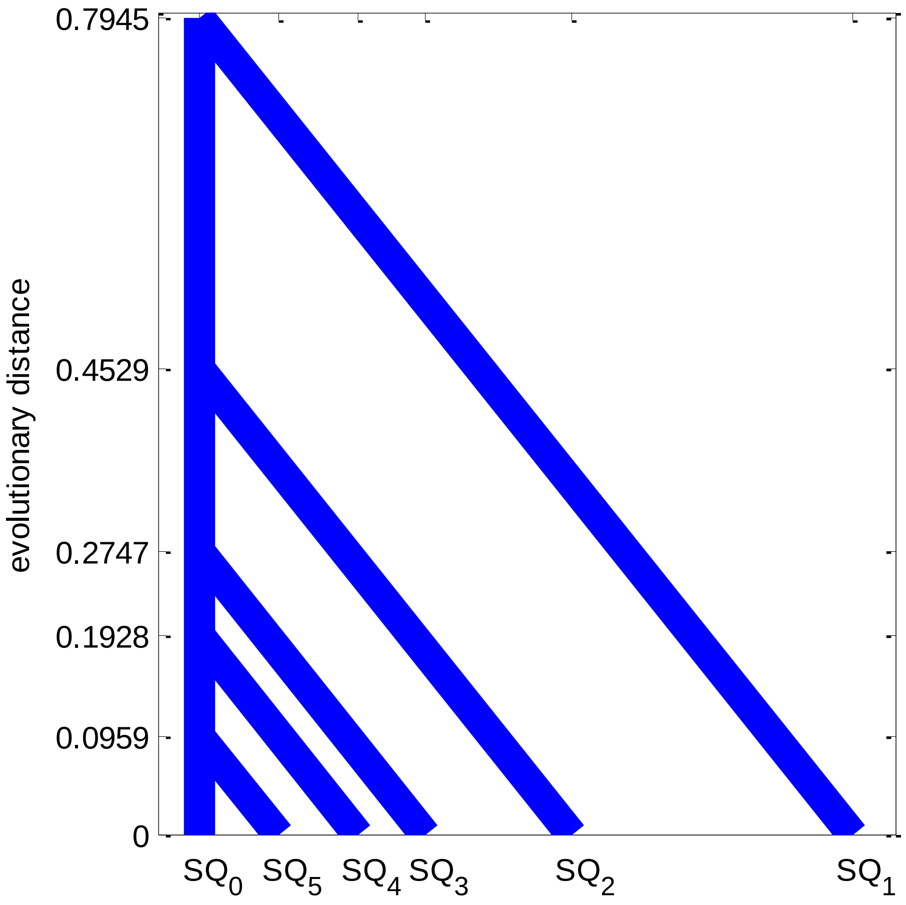
AGCAACCGTGCCCATCGAAG...

AGTACCTGCGGCCATCGAAG...

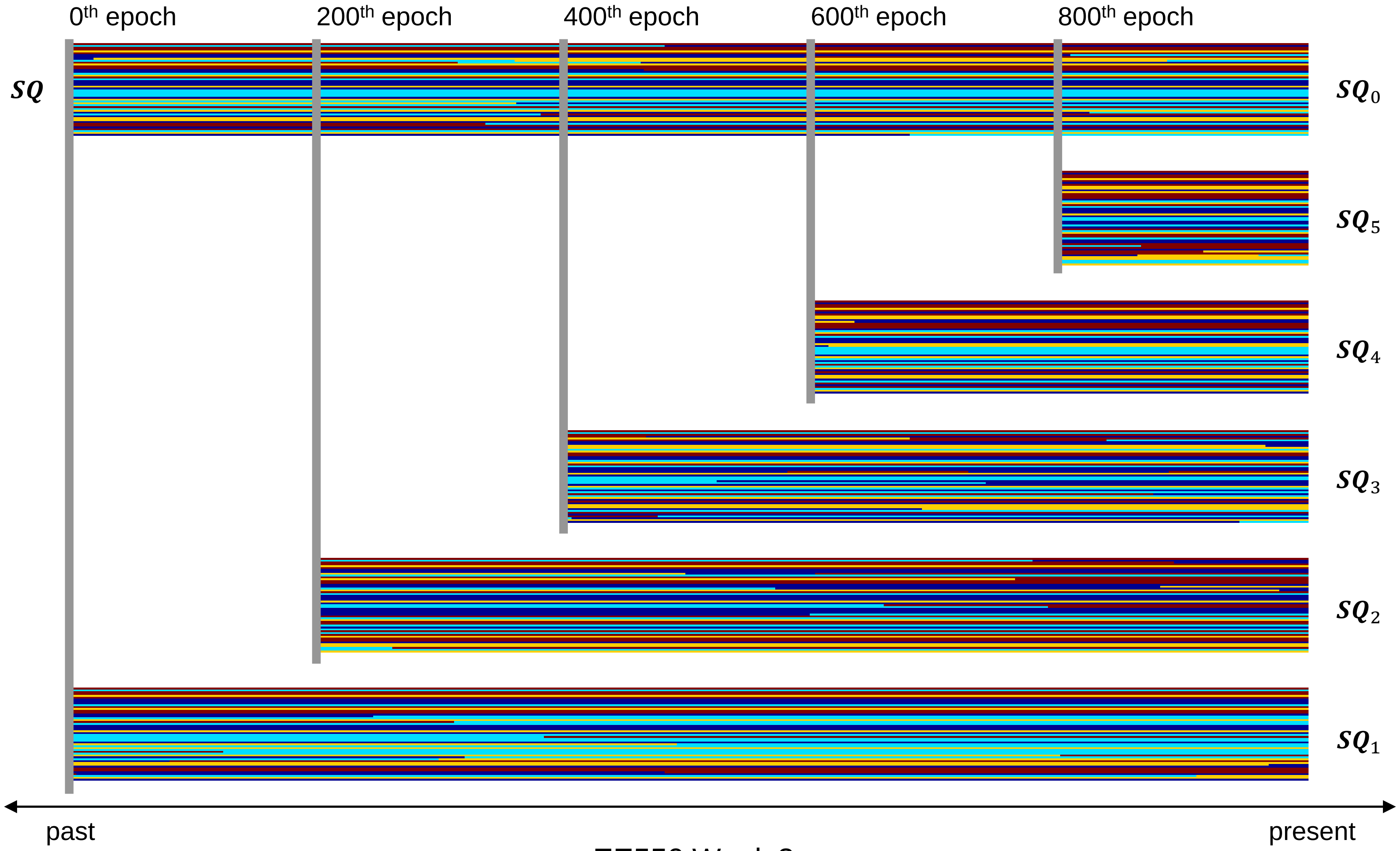
$\boldsymbol{SQ_5}$

Example: Evolutionary Distances

- Sequence distances:
  - $D_{0,1} = 0.4900 \Rightarrow d_{0,1} = 0.7945$
    ```
    AGTACCCGGGGCCATCGAAG…
     1 1    11 11      1…
    ATTTCCCGTCGAGATCGAAT…
    ```
  - $D_{0,2} = 0.3400 \Rightarrow d_{0,2} = 0.4529$
    ```
    AGTACCCGGGGCCATCGAAG…
     1      11  1 11    …
    ATTACCCGTTGCGAGGGAAG…
    ```
  - $D_{0,3} = 0.2300 \Rightarrow d_{0,3} = 0.2747$
    ```
    AGTACCCGGGGCCATCGAAG…
         1  1   1     1 …
    AGTACACGTGGCAATCGAGG…
    ```
  - $D_{0,4} = 0.1700 \Rightarrow d_{0,4} = 0.1928$
    ```
    AGTACCCGGGGCCATCGAAG…
      1 1   1 1         …
    AGCAACCGTGCCCATCGAAG…
    ```
  - $D_{0,5} = 0.0900 \Rightarrow d_{0,5} = 0.0959$
    ```
    AGTACCCGGGGCCATCGAAG…
          1 1           …
    AGTACCTGCGGCCATCGAAG…
    ```



Repeat Example: Sequence Data

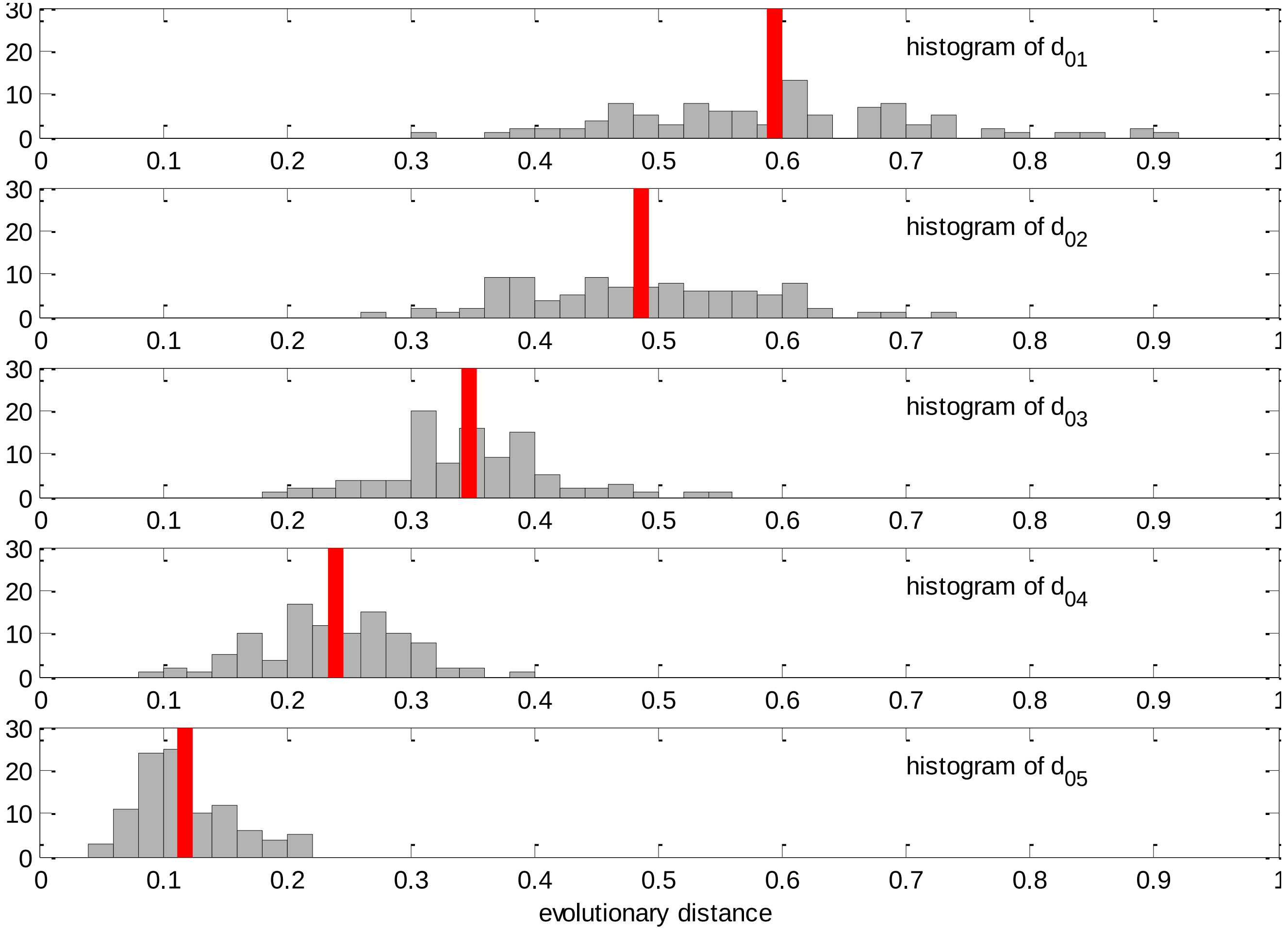


Repeat Example: Evolutionary Distances

- Distances:
  - $D_{0,1} = 0.4400 \Rightarrow d_{0,1} = 0.6626$
  - $D_{0,2} = 0.3000 \Rightarrow d_{0,2} = 0.3831$
  - $D_{0,3} = 0.3000 \Rightarrow d_{0,3} = 0.3831$
  - $D_{0,4} = 0.1500 \Rightarrow d_{0,4} = 0.1674$
  - $D_{0,5} = 0.0800 \Rightarrow d_{0,5} = 0.0846$
- Remark:
  - Even though the experiment setup is exactly the same, the distances vary
    - Sequence evolution is a stochastic process
  - The variation even produces a rather strange and quite disagreeable result:

$D_{0,2} = D_{0,3}$ providing $d_{0,2} = d_{0,3}$!!!

Example: Variability in Computed Evolutionary Distances



Remarks

- Models of nucleic acid sequence evolution links the sequence differences to evolutionary distances
- The parameters of these models are fitted to the available data to capture reality as much as possible
  - More sophisticated models better fit the available data
  - With better fits, the risk of losing general validity increases
- The viability of these models depends on the validity of the premises on the given application data
  - Assumptions may not hold
- The estimated evolutionary distances, however, are subject to estimation errors
  - These errors may switch the order of evolutionary siblinghood
- The extent of errors are directly proportional to the expected evolutionary distances
  - For sequences that are similar, the expected error is small
  - For sequences that are significantly different, the errors are large