

Substitution Models and the Phylogenetic Assumptions

Vivek Jayaswal
Lars S. Jermiin

COMMONWEALTH OF AUSTRALIA

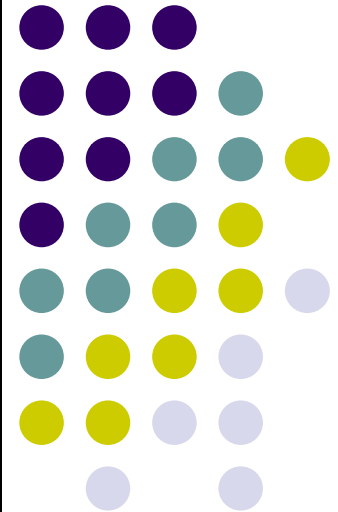
Copyright Regulation

WARNING

This material has been reproduced and communicated to you by or on
be half of the University of Sydney pursuant to Part VB of the
Copyright Act 1968 (the Act).

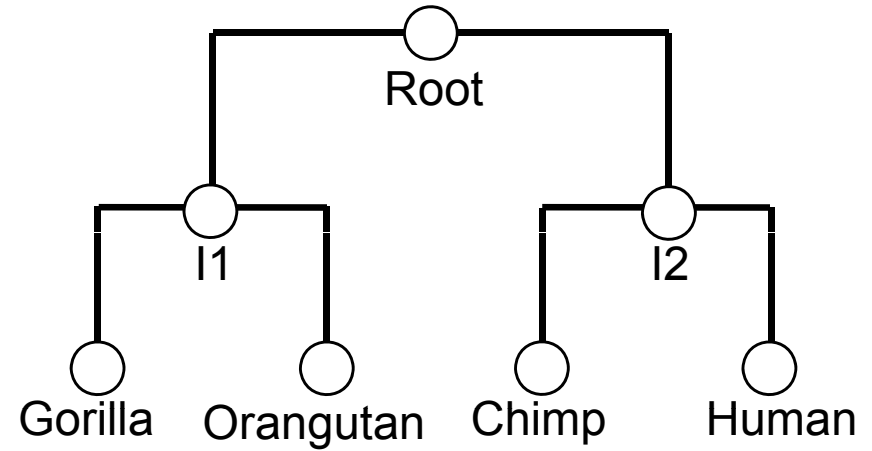
The material in this communication may be subject to copyright under
the Act. Any further reproduction or communication of this material by
you may be the subject of copyright protection under the Act.

Do not remove this notice.

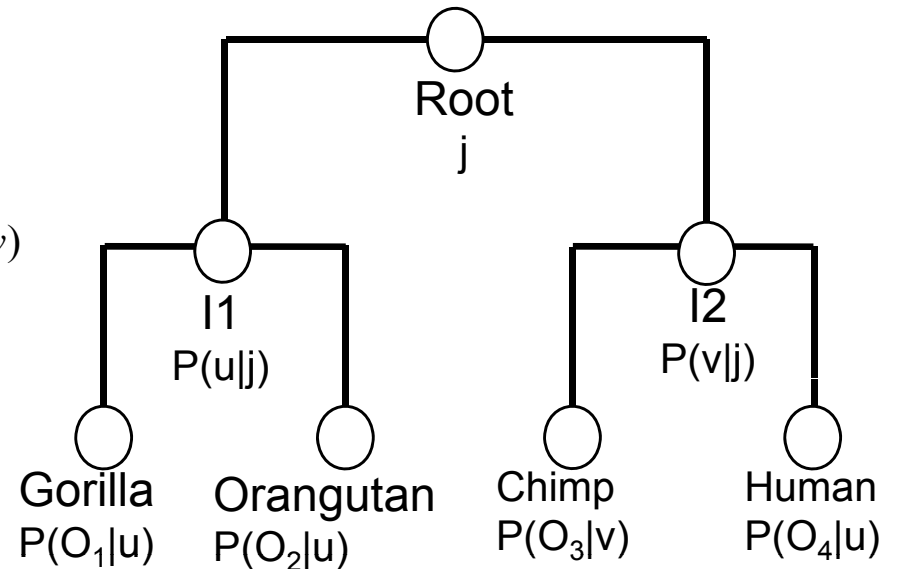


Why do we need substitution models?

	1	2	3	4	5	...	N
Human	A	T	C	G	A	...	C
Chimp	A	G	C	A	A	...	C
Gorilla						
Orangutan						



$$L_i = \sum_{j=1}^4 f_j \sum_{u=1}^4 P(u | j) \sum_{v=1}^4 P(v | j) P(O_1 | u) P(O_2 | u) P(O_3 | v) P(O_4 | v)$$

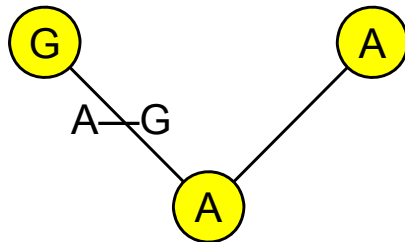


Topics

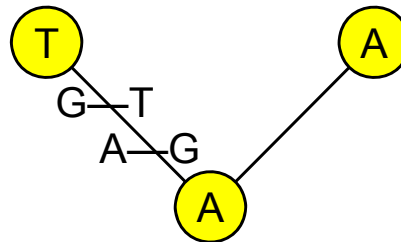
- how correction for multiple substitution are done
- some of the phylogenetic assumptions
- how we may evaluate phylogenetic assumptions
- an example involving bacterial DNA

Substitutions at a Single Site

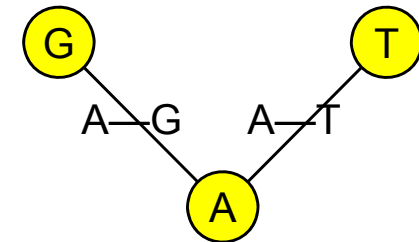
Single Substitution



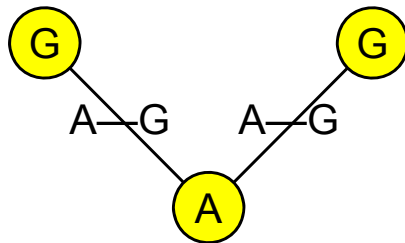
Multiple Substitution



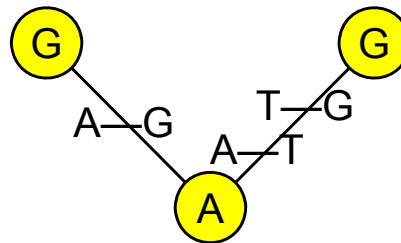
Coincidental Substitution



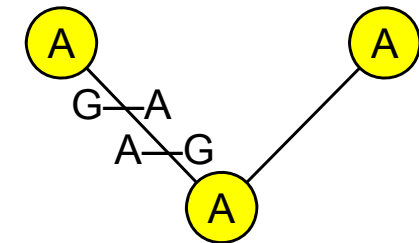
Parallel Substitution



Convergent Substitution



Back Substitution

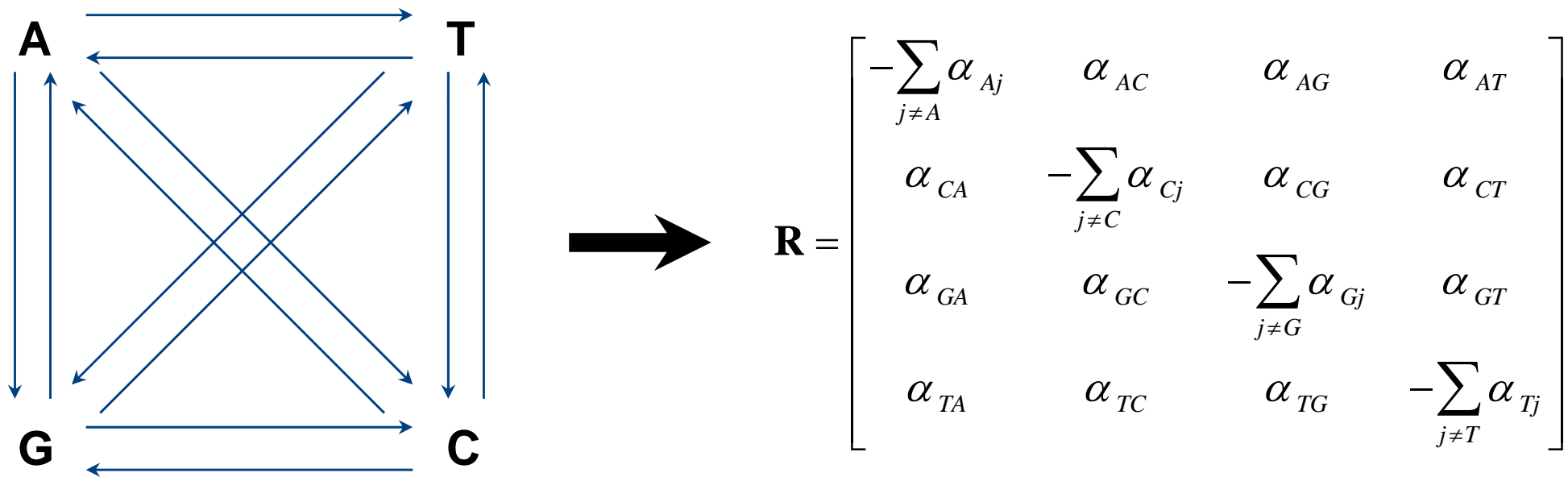


Note —

- Every substitution overwrites the evidence of a past state, which leads to an **erosion of the historical signal**

Modeling Nucleotide Substitutions

Consider the evolution at a given site in terms of **conditional rates-of-change** from nucleotide i to nucleotide j



Note —

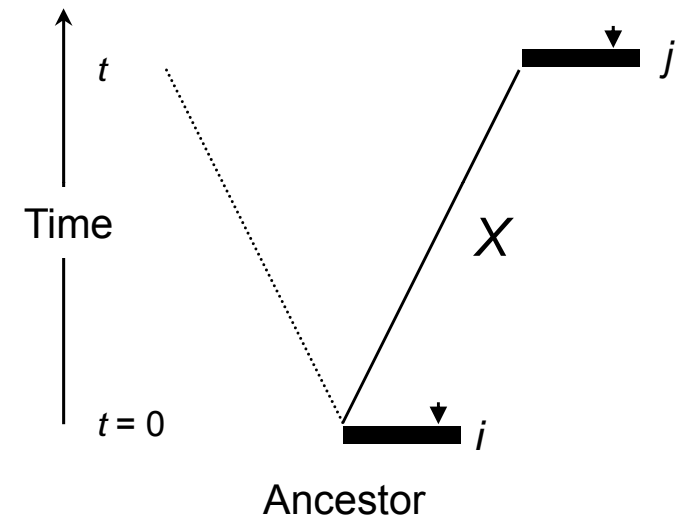
- Here α_{ij} is the **conditional rate-of-change** from nucleotide i to nucleotide j in \mathbf{R} — the ‘rate matrix’ — and \mathbf{R} is the most **general Markov model** for DNA

Modelling a Site in one Sequence

Consider **Markov process**, X , that results in nucleotide i being converted to nucleotide j over time t —

$$P_{ij}(t) = P[X(t) = j \mid X(0) = i]$$

If the **rates** of change are constant, then $P(t) = e^{\mathbf{R}t}$



In **matrix notation**, this is

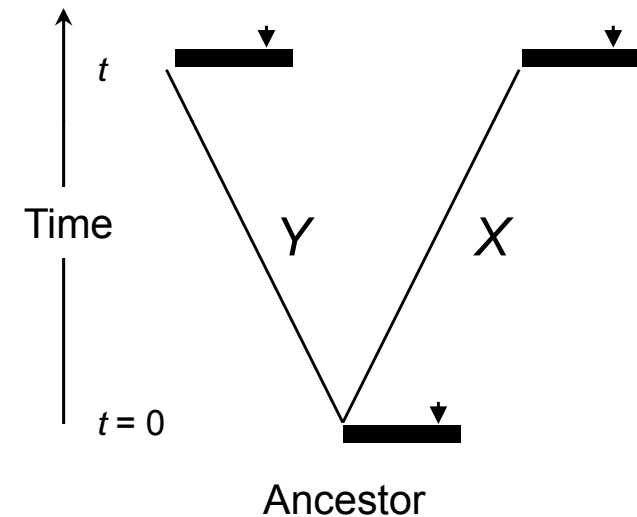


$$\begin{aligned} \mathbf{P}(t) &= \mathbf{I} + \mathbf{R}t + \frac{(\mathbf{R}t)^2}{2!} + \frac{(\mathbf{R}t)^3}{3!} + \dots \\ &= \sum_{k=0}^{\infty} \frac{(\mathbf{R}t)^k}{k!} \end{aligned}$$

Modelling a Site in two Sequences

Consider two **Markov processes**, X and Y and the following scenario —

$$F_{ij}(t) = P[X(t) = i, Y(t) = j \mid X(0) = Y(0)]$$



In **matrix notation**, this is



$$\mathbf{F}(t) = (\mathbf{P}^X(t))^T \mathbf{F}(0) \mathbf{P}^Y(t)$$

Take-home Message #1

- The substitution model, \mathbf{R} , is an integral part of the transition function; it is used to estimate the probability of the present states, given \mathbf{R} and t

Rate matrix revisited

$$\mathbf{R}_i = \begin{bmatrix} -\sum_{j \neq A} r_{Aj} & r_{AC} & r_{AG} & r_{AT} \\ r_{CA} & -\sum_{j \neq C} r_{Cj} & r_{CG} & r_{CT} \\ r_{GA} & r_{GC} & -\sum_{j \neq G} r_{Gj} & r_{GT} \\ r_{TA} & r_{TC} & r_{TG} & -\sum_{j \neq T} r_{Tj} \end{bmatrix}$$

- Typically simplified forms of this rate matrix are used
- These matrices belong to the GTR-family of models and can be represented as

$$\mathbf{R} = \begin{bmatrix} - & \mathbf{r}_{AC} & \mathbf{r}_{AG} & \mathbf{r}_{AT} \\ \mathbf{r}_{AC} & - & \mathbf{r}_{CG} & \mathbf{r}_{CT} \\ \mathbf{r}_{AG} & \mathbf{r}_{CG} & - & \mathbf{r}_{GT} \\ \mathbf{r}_{AT} & \mathbf{r}_{CT} & \mathbf{r}_{GT} & - \end{bmatrix} \begin{bmatrix} \boldsymbol{\pi}_A \\ \boldsymbol{\pi}_C \\ \boldsymbol{\pi}_G \\ \boldsymbol{\pi}_T \end{bmatrix}$$

Commonly-used Markov Models

Jukes & Cantor (1969)

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



Assumptions

1. One rate (α)
2. Uniform nucleotide content

Kimura (1980)

$$\mathbf{R} = \begin{bmatrix} -(2\beta + \alpha) & \beta & \alpha & \beta \\ \beta & -(2\beta + \alpha) & \beta & \alpha \\ \alpha & \beta & -(2\beta + \alpha) & \beta \\ \beta & \alpha & \beta & -(2\beta + \alpha) \end{bmatrix}$$



1. Two rates (α and β)
2. Uniform nucleotide content

Hasegawa, Kishino & Yano (1985)

$$\mathbf{R} = \begin{bmatrix} -(\beta\pi_Y + \alpha\pi_G) & \beta\pi_C & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & -(\beta\pi_R + \alpha\pi_T) & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & -(\beta\pi_Y + \alpha\pi_A) & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & -(\beta\pi_R + \alpha\pi_C) \end{bmatrix}$$



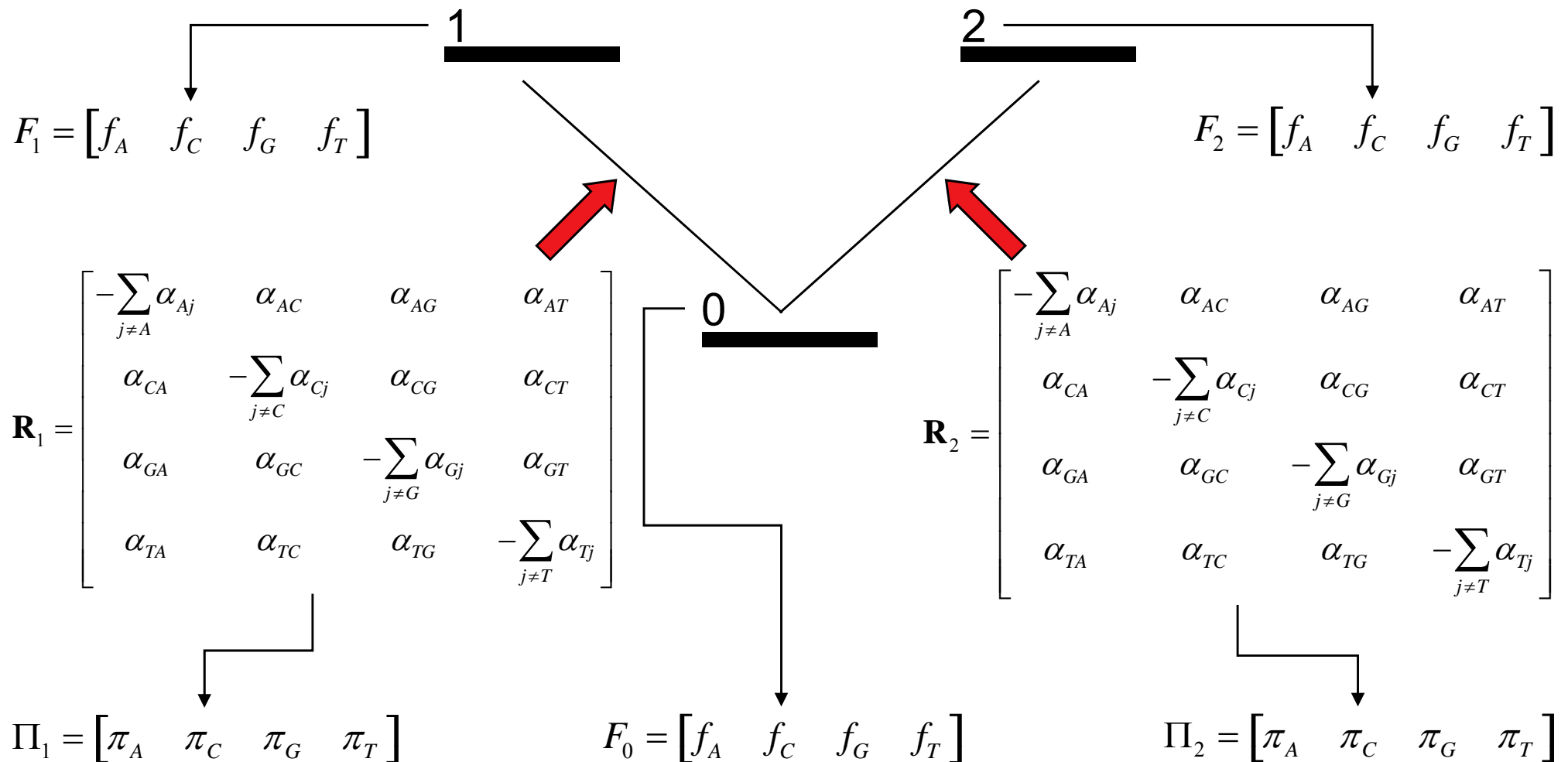
1. Two rates (α and β)
2. Non-uniform nucleotide content

The Phylogenetic Assumptions

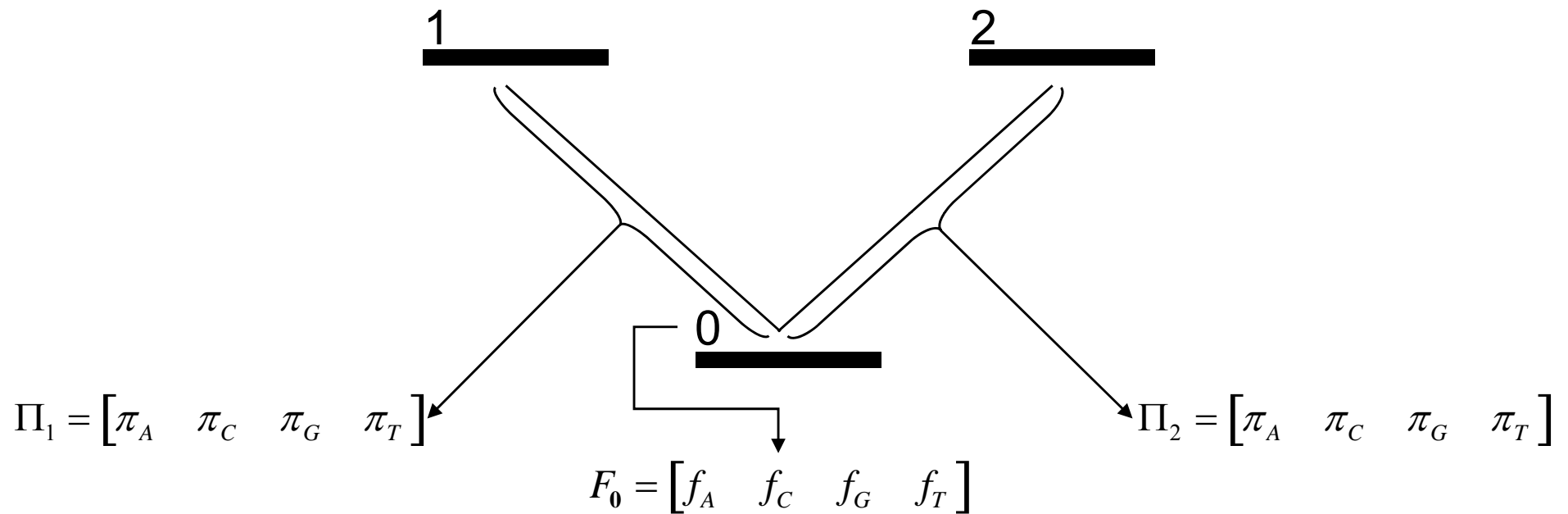
- Given an alignment of nucleotides, phylogenetic methods commonly assume that the sites have evolved under
 - **stationary** and **reversible** conditions
 - **homogeneous** conditions
 - **independent** and **identical** conditions

Phylogenetic Assumptions

Consider the following scenario...



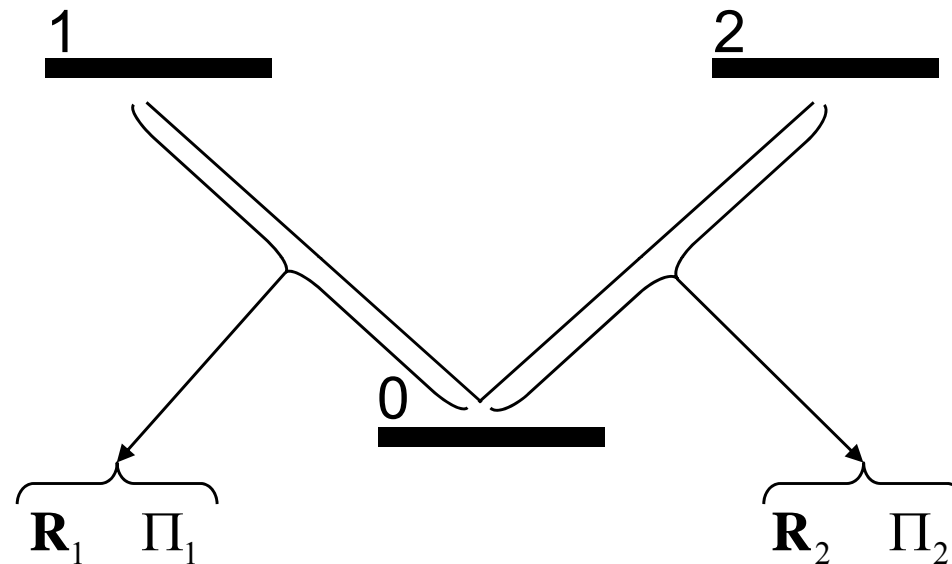
The Stationary Condition



Note —

- The **stationary condition** is met if $F_0 = \Pi_1 = \Pi_2$, implying that the marginal distributions of \mathbf{R}_1 and \mathbf{R}_2 are the same, even though \mathbf{R}_1 and \mathbf{R}_2 may differ!

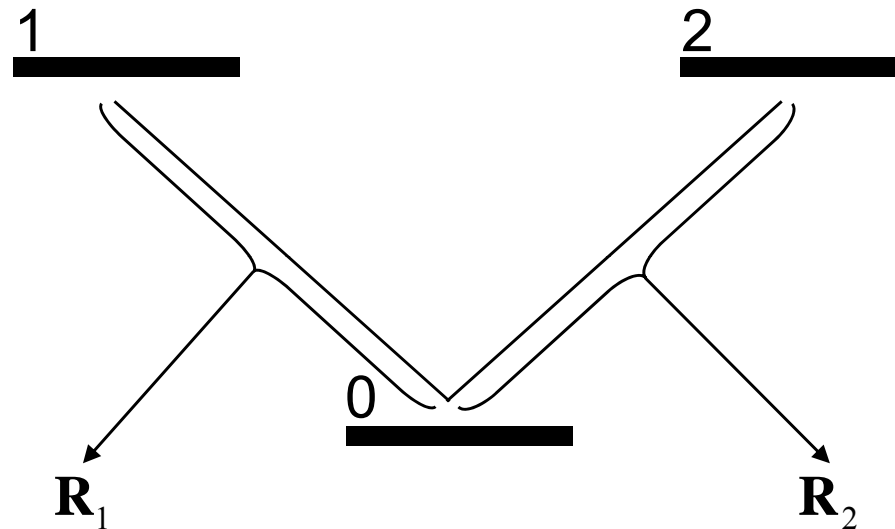
The Reversible Condition



Notes —

- If the process is stationary, then $\Pi_1 = \Pi_2$
- Moreover, if $\pi_i \mathbf{R}_{ij} = \pi_j \mathbf{R}_{ji}$ for all i and j , then the process is reversible

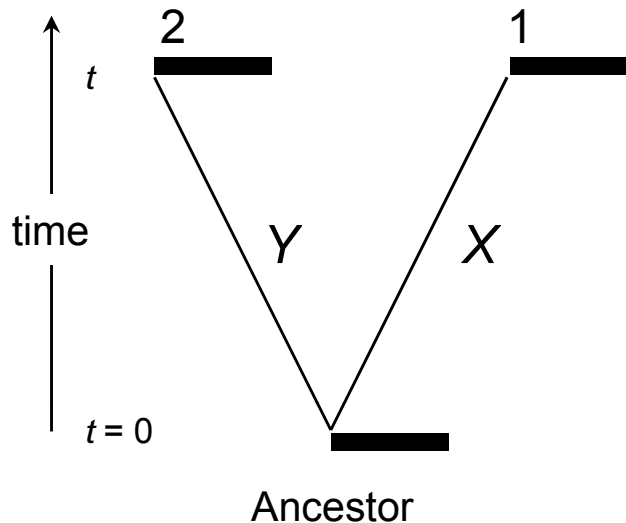
The Homogeneous Condition



Notes —

- The **homogeneous condition** is met for the Markov processes, R_1 and R_2 , if $R_1 = R_2$
- If the homogeneous condition is met, then $\Pi_1 = \Pi_2$ — however, non-stationary, and therefore non-reversible, conditions may still prevail (i.e., if $F_0 \neq \Pi_1 = \Pi_2$)

IID condition



```
Seq_1  ACGTGTCCATGATTA...
        ||
        Rx Rx ...
        ||
Seq_2  ACCTGCCCAAGATAA...
```

Notes —

For computational reasons, it is convenient to assume that —

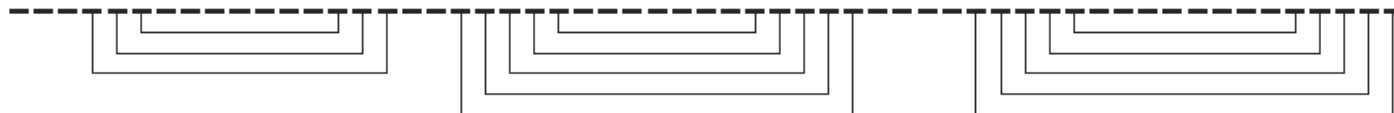
- Sites in the sequence have evolved **independently**
- Sites in the sequence have evolved **under the same model** ($R_x = R_y$)

Rate Heterogeneity Across Sites

RNA-coding Genes

Gene GAACTTGATTTAAAAGCCTATGTTTTGAAAACATAATAAGAAATATAAATTTTTCT
Unit -----

Gene GAACTTGATTTAAAAGCCTATGTTTTGAAAACATAATAAGAAATATAAATTTTTCT
Unit -----
Category 2223332222223332233332211122333322223333222222333333

Gene GAACTTGATTTAAAAGCCTATGTTTTGAAAACATAATAAGAAATATAAATTTTTCT
Unit -----

Category 2223332222223332233332211122333322223333222222333333

Take-home Message #2

- Phylogenetic analyses require the users to make certain assumptions about the data before these are investigated in detail

Questions...

- Are these phylogenetic assumptions realistic?
- How can we assess whether the phylogenetic assumptions are met by the data?

Answer...

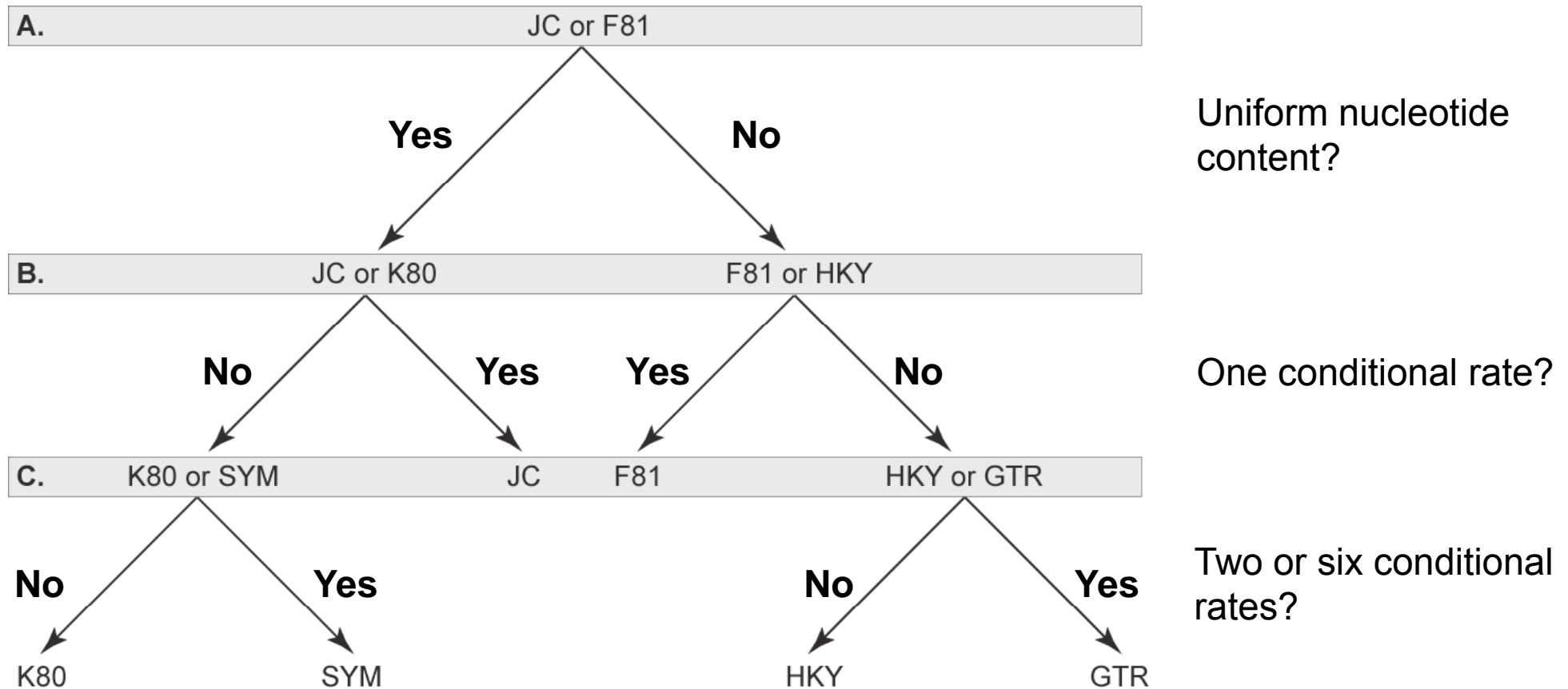
- Inspecting the data — before or after inferring the phylogeny — increases the chance of finding out what might have taken place during the evolution

Considering the IID Condition

- **Visual inspection** of the alignment might show whether some regions evolved faster
- Assume **rate-heterogeneity across some sites**, and then use phylogenetic methods that account for this

Hierarchical Likelihood-Ratio Test

Consider the following decision tree...



Take-home Message #3

- The **likelihood-ratio test** can be used identify a suitable substitution model for a given data set

Testing assumptions prior to modelling

- $n \times \mathbf{F}(t)$ — expected divergence matrix
- $\mathbf{N}(t)$ — observed divergence matrix

Examples:

$$\mathbf{N}(0) = \begin{bmatrix} 294 & 0 & 0 & 0 \\ 0 & 372 & 0 & 0 \\ 0 & 0 & 829 & 0 \\ 0 & 0 & 0 & 655 \end{bmatrix}$$

$$\mathbf{N}(t) = \begin{bmatrix} 244 & 31 & 8 & 11 \\ 28 & 321 & 14 & 9 \\ 11 & 13 & 801 & 4 \\ 14 & 10 & 3 & 628 \end{bmatrix}$$

Matched-pairs Tests of Symmetry

Seq 1 AGACTAGGTCTTGTATAGACTAATGTTTCACAGTTTTTTAACTTTGTCAATGGA...
 Seq 2 AGACGAGGTCGTGTATGGCCTCGTGAGCACGGGTTGTTCACTCCGCCAACGGT...

	A	C	G	T	Σ_2
A	5	4	7	1	17
C	0	7	2	0	9
G	0	1	5	0	6
T	1	4	5	8	18
Σ					
1	6	16	19	9	

Note:

- These tests statistics are asymptotically χ^2 -distributed on ν degrees of freedom

Test of Symmetry

$$X_{\text{Bowker}}^2 = \sum_{i < j} \frac{(x_{ij} - x_{ji})^2}{x_{ij} + x_{ji}}$$

Test of Marginal Symmetry

$$X_{\text{Stuart}}^2 = \mathbf{D}\mathbf{V}^{-1}\mathbf{D}^T,$$

$$\mathbf{D}_{ij} = x_{i\bullet} - x_{\bullet i},$$

$$\mathbf{V} = \text{covariance matrix of } \mathbf{D}$$

Bacterial 16S rDNA Sequences

Ribosomal RNA from five bacteria was compared using the matched-pairs test of homogeneity

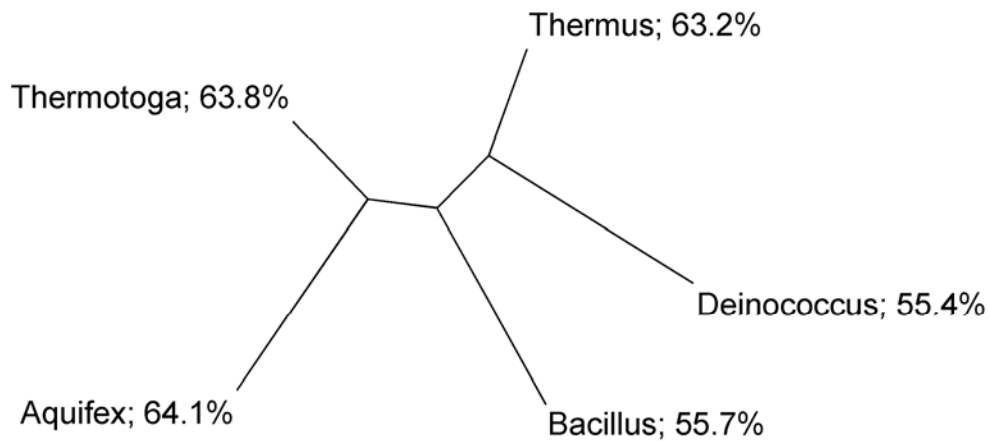
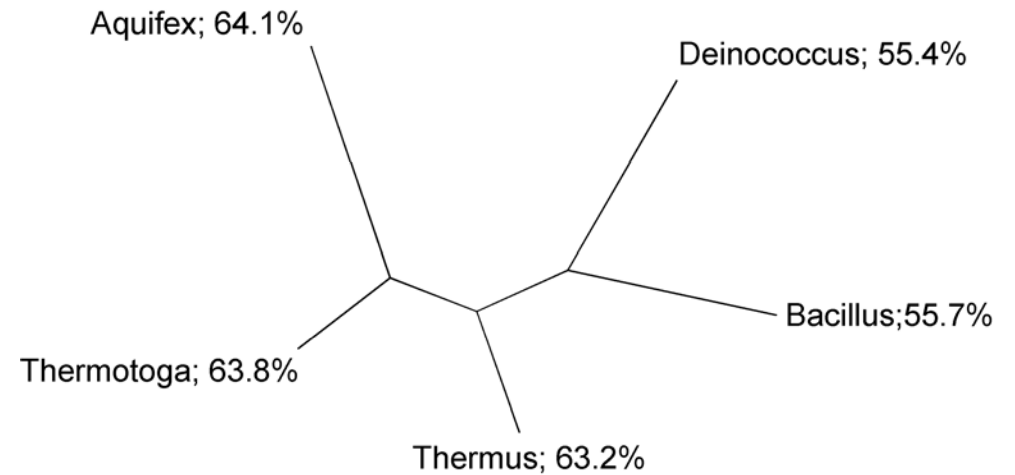
Probabilities	<i>Thermotoga</i>	<i>Bacillus</i>	<i>Deinococcus</i>	<i>Thermus</i>
<i>Aquifex</i>	1.32 10 ⁻⁰¹	1.79 10 ⁻¹¹	3.45 10 ⁻¹⁰	5.09 10 ⁻⁰¹
<i>Thermotoga</i>		2.64 10 ⁻¹²	6.69 10 ⁻¹²	4.15 10 ⁻⁰¹
<i>Bacillus</i>			9.95 10 ⁻⁰¹	3.64 10 ⁻⁰⁹
<i>Deinococcus</i>				5.99 10 ⁻¹¹

Note —

It is highly unlike that these data have evolved under homogeneous conditions, implying that it would be unwise to use a time-reversible Markov model

Phylogeny of Bacterial Ribosomal RNA

Markov model: **GTR**



Markov model: **General**

Take-home Message #4

- It is **important** to consider the substitution models carefully when using them in phylogenetic studies

Suggested Literature

RDM Page, EC Holmes (1998), **Molecular Evolution**.

- Chapter 5 (Sections 5.2, 5.3) — important reading

W-H Li (1997), **Molecular Evolution**.

- Chapter 3 (pp. 59-78) — contains descriptions that are better than those in Page & Holmes 1998 — important reading

D Posada, KA Crandall, 1998. MODELTEST: testing the model of DNA substitution. **Bioinformatics** 14, 817-818 — useful reading

LS Jermin *et al.* (2008). Phylogenetic model Evaluation. Pp 331-363. In **Bioinformatics - Volume I: Data, Sequences Analysis and Evolution** (Ed. Keith J), Humana Press, Totowa, NJ. [2008] — important reading