# GPU-CAPABLE SOFTWARE FOR STATISTICAL PHYLOGENETICS

Will Landau, Matt Simpson, Prof. Jarad Niemi

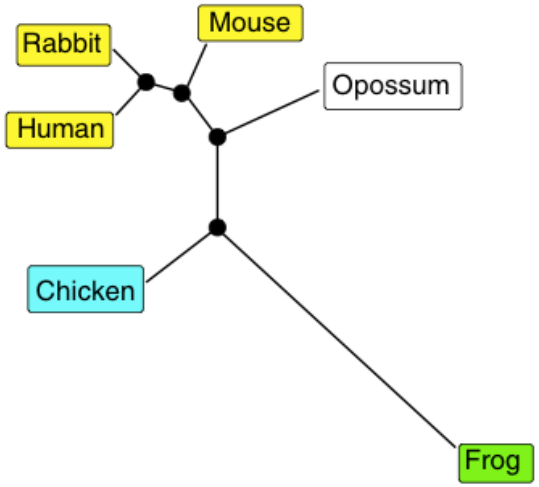
# INTRODUCTION TO PHYLOGENETICS

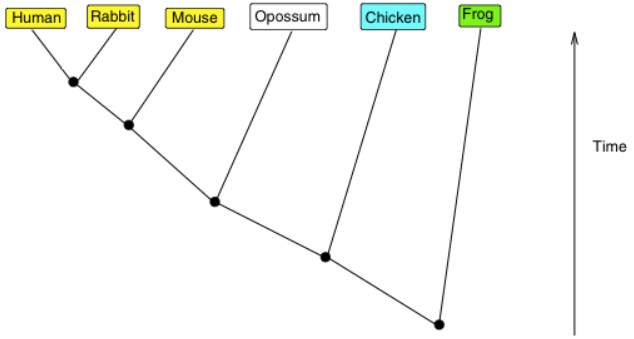
**Phylogenetics:** The study of evolutionary relationships among different species or strains of organisms.

Given a set of species, the goal is to construct a model that describes:

1. Lineage and ancestry.
2. The relative degree of genetic similarity between and among the different organisms.

# PHYLOGENETIC TREES



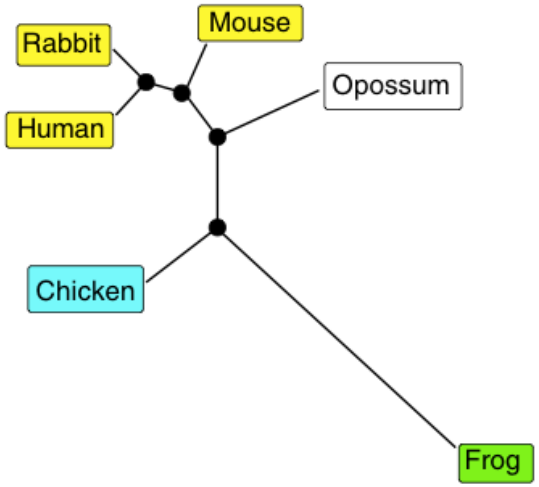




- The leaf nodes are the known, contemporary species that we care about
- The inner nodes are hidden (possibly unknown) ancestors
- An edge identifies a pair of nodes in which one member evolved directly from the other.

In general, a **phylogenetic tree** is an acyclic graph. The following properties have explicit phylogenetic meaning.

- **topology**: the branching structure: i.e., which nodes are connected to which. Topology depicts ancestry.
- **branch lengths**: how long each edge is. Branch lengths are proportional to genetic distance as measured in phylogenetic time, $\lambda \cdot t$, where $\lambda$ is the genetic mutation rate and $t$ is physical time.

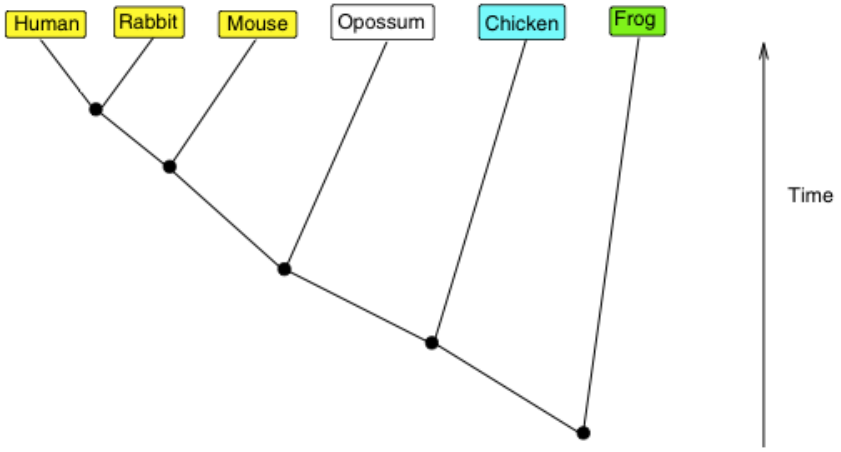
# UNROOTED TREES




An **unrooted tree** is an UNDIRECTED acyclic graph that indicates no common ancestor.

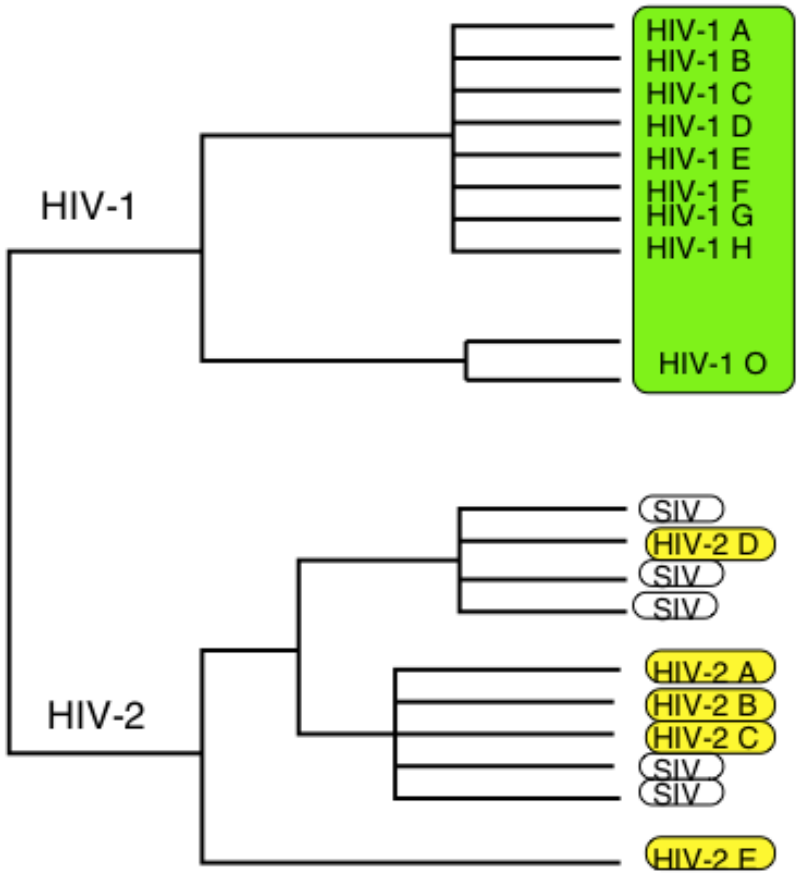
We add the additional assumption that the leaves evolved from the inner nodes, but otherwise, we don't know what evolved from what.

# UNROOTED TREES

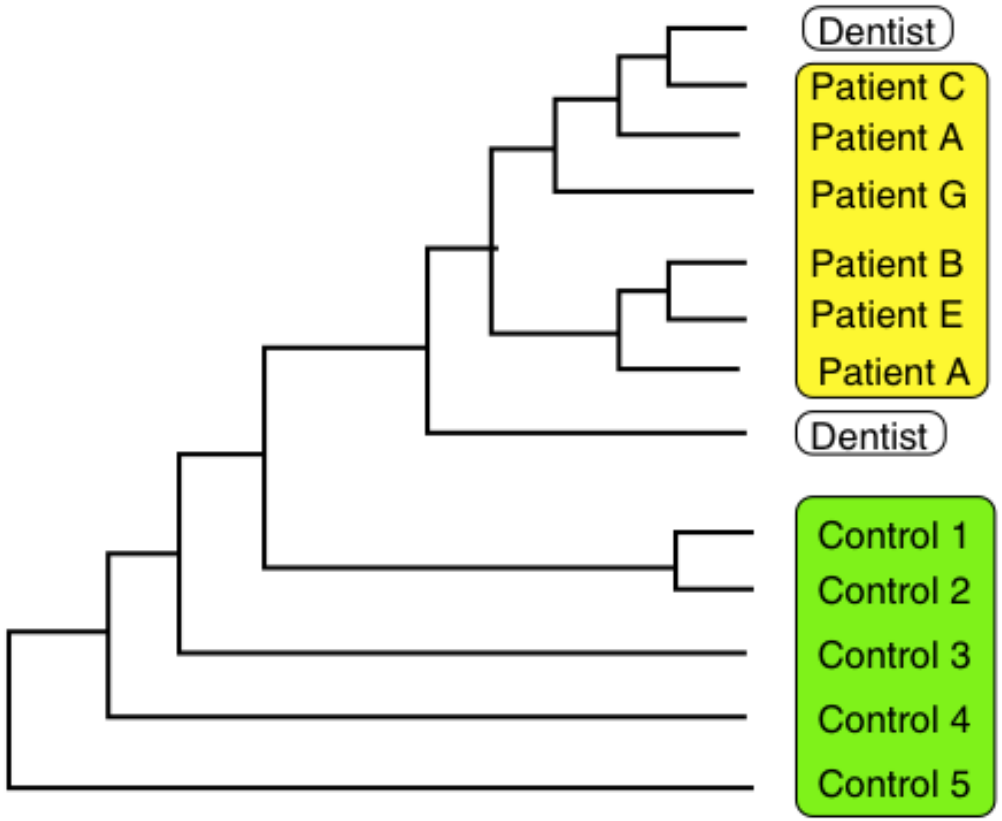



An **rooted tree** is a DIRECTED acyclic graph that indicates no common ancestor.

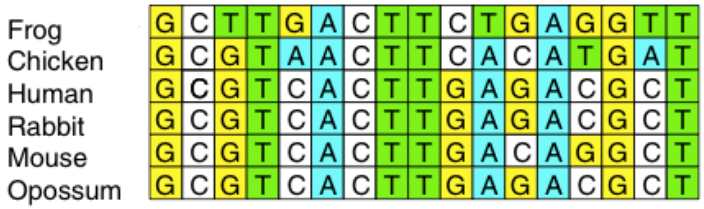
The time axis indicates the direction of the edges (the direction of evolution).

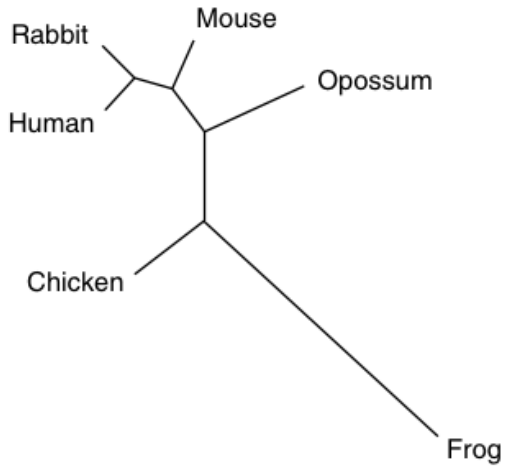



Some phylogenetic trees are displayed as cladograms like this one. The horizontal edge lengths convey phylogenetic distance, and the vertical lines are meaningless.




This one revealed that a Florida dentist infected several of his patients with HIV around 1990.



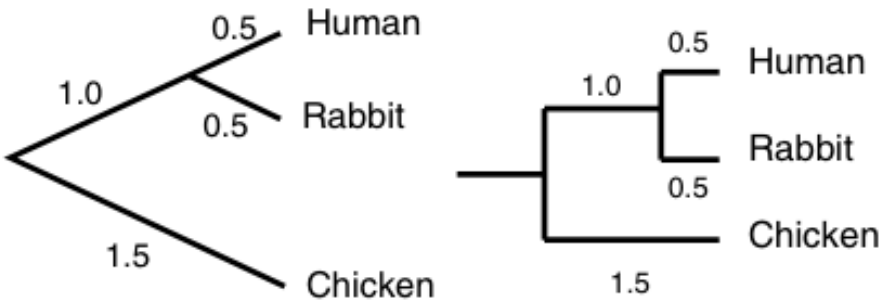

Topologies and branch lengths are inferred by sequencing DNA from each of the contemporary organisms and determining the degree of similarity from DNA sequence alignment.

Human ... T G T **A** T C G C T C ...
Rabbit ... T G T **G** T C G C T C ...

Human ... **T** G T **A** T C G **C** T C ...
Chicken ... **A** G T **C** T C G **T** T C ...

Rabbit ... **T** G T **G** T C G **C** T C ...
Chicken ... **A** G T **C** T C G **T** T C ...

| | Rabbit | Chicken |
|---|---|---|
| Human | 1 | 3 |
| Rabbit | | 3 |




Use of genetic distances: one simple and naive way to put together a phylogenetic tree.

# WAYS TO INFER PHYLOGENETIC TREES FROM GENOMIC DATA

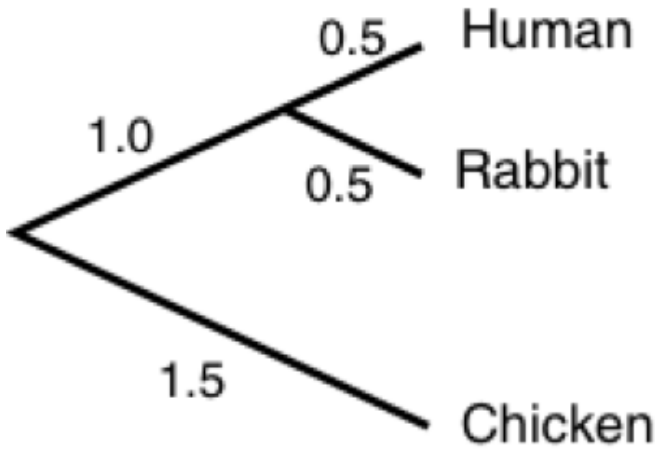
- Use of genetic distances
- UPGMA Clustering (Unweighted Pair Group Method using Arithmetic averages)
- Clustering via neighbor joining
- Parsimony
- **MAXIMUM LIKELIHOOD USING CONTINUOUS TIME MARKOV CHAINS TO MODEL GENETIC SEQUENCE MUTATION OVER TIME**

# RECAP: WE CARE ABOUT SEQUENCE ALIGNMENT FOR CONSTRUCTING TREES

Human ... T G T **A** T C G C T C ...
Rabbit ... T G T **G** T C G C T C ...

Human ... **T** G T **A** T C G **C** T C ...
Chicken ... **A** G T **C** T C G **T** T C ...

Rabbit ... **T** G T **G** T C G **C** T C ...
Chicken ... **A** G T **C** T C G **T** T C ...

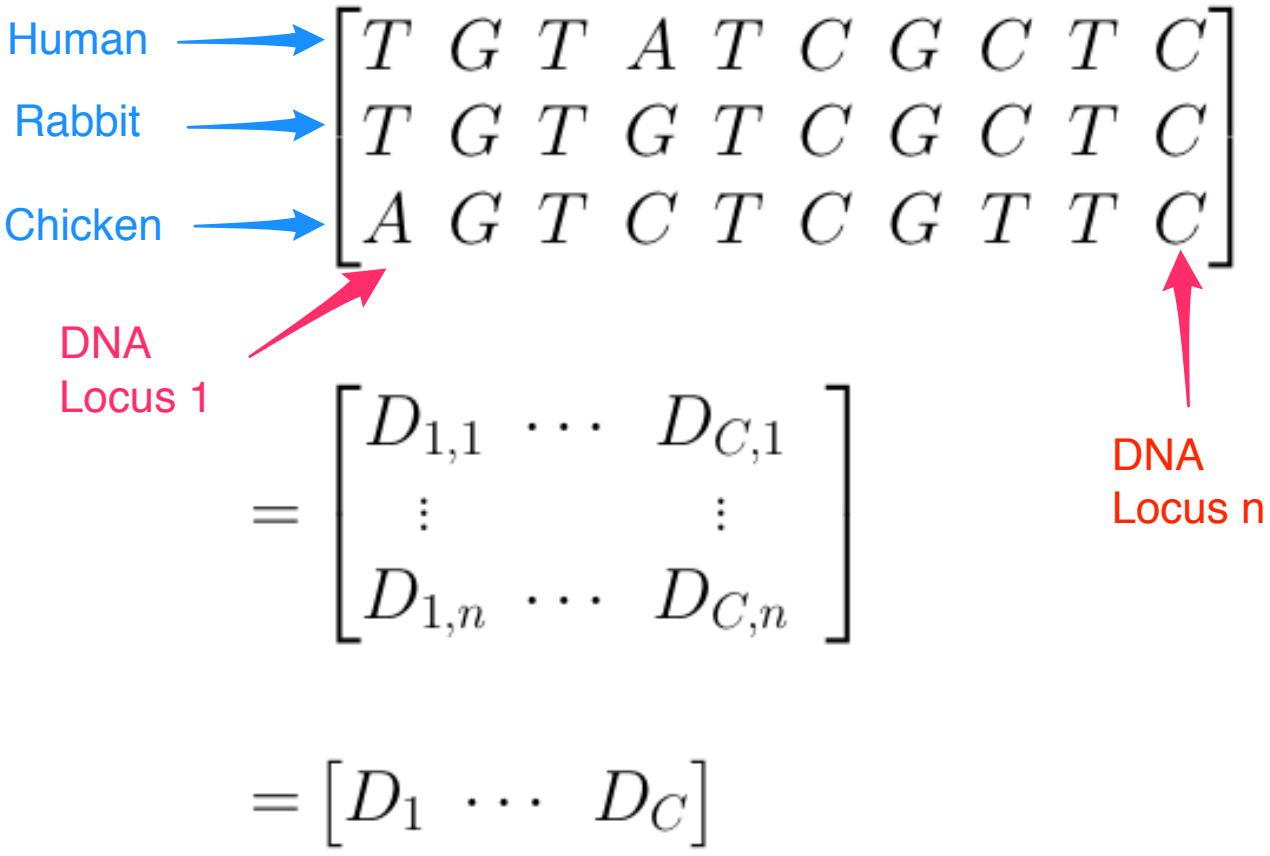


# BY WHAT PROCESS DID THE RABBIT AND THE CHICKEN EACH DESCEND FROM THEIR MOST RECENT COMMON ANCESTOR?

Rabbit ... **T** G T **G** T C G **C** T C ...
Chicken ... **A** G T **C** T C G **T** T C ...

Use a continuous time Markov process to model the genetic mutations from ancestor to rabbit and ancestor to chicken.

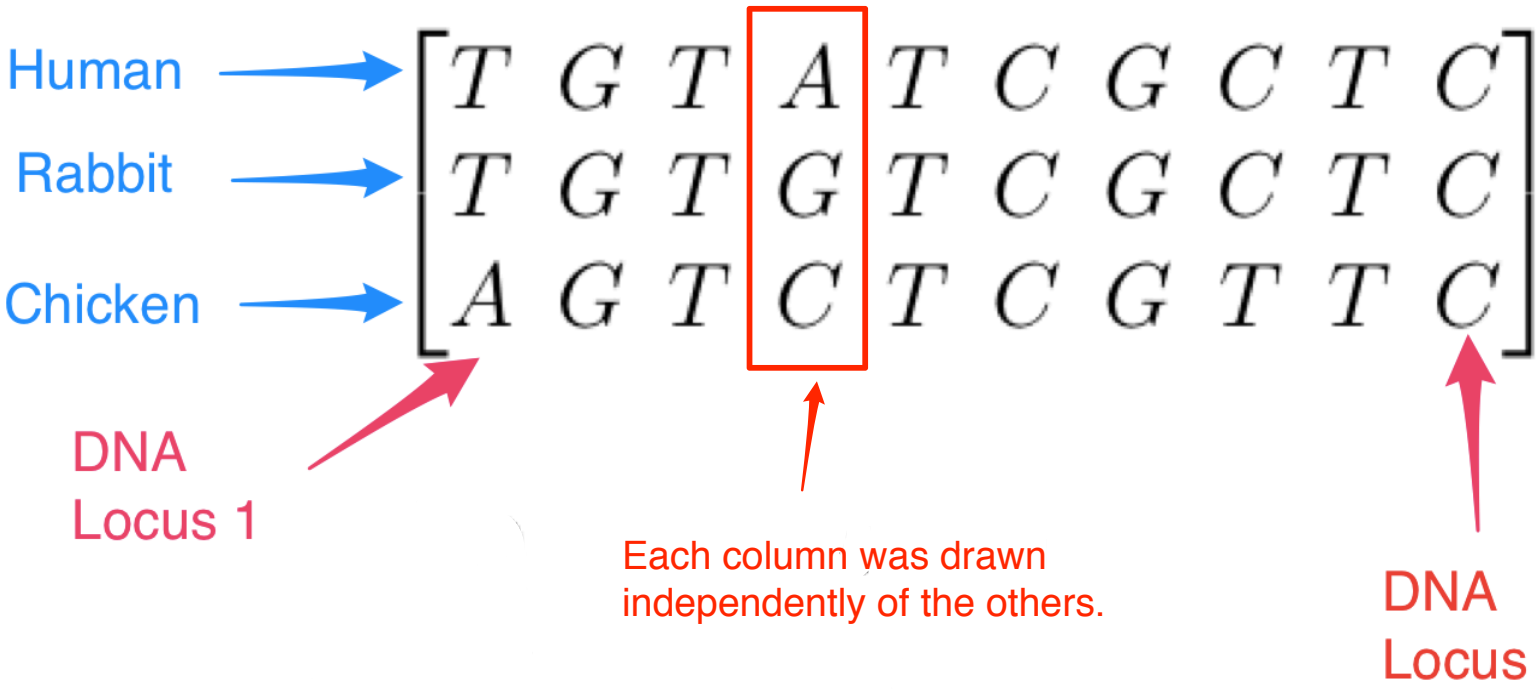
Sequence alignment data $\mathbf{D} =$




$$\begin{bmatrix} T & G & T & A & T & C & G & C & T & C \\ T & G & T & G & T & C & G & C & T & C \\ A & G & T & C & T & C & G & T & T & C \end{bmatrix}$$

$$= \begin{bmatrix} D_{1,1} & \cdots & D_{C,1} \\ \vdots & & \vdots \\ D_{1,n} & \cdots & D_{C,n} \end{bmatrix}$$

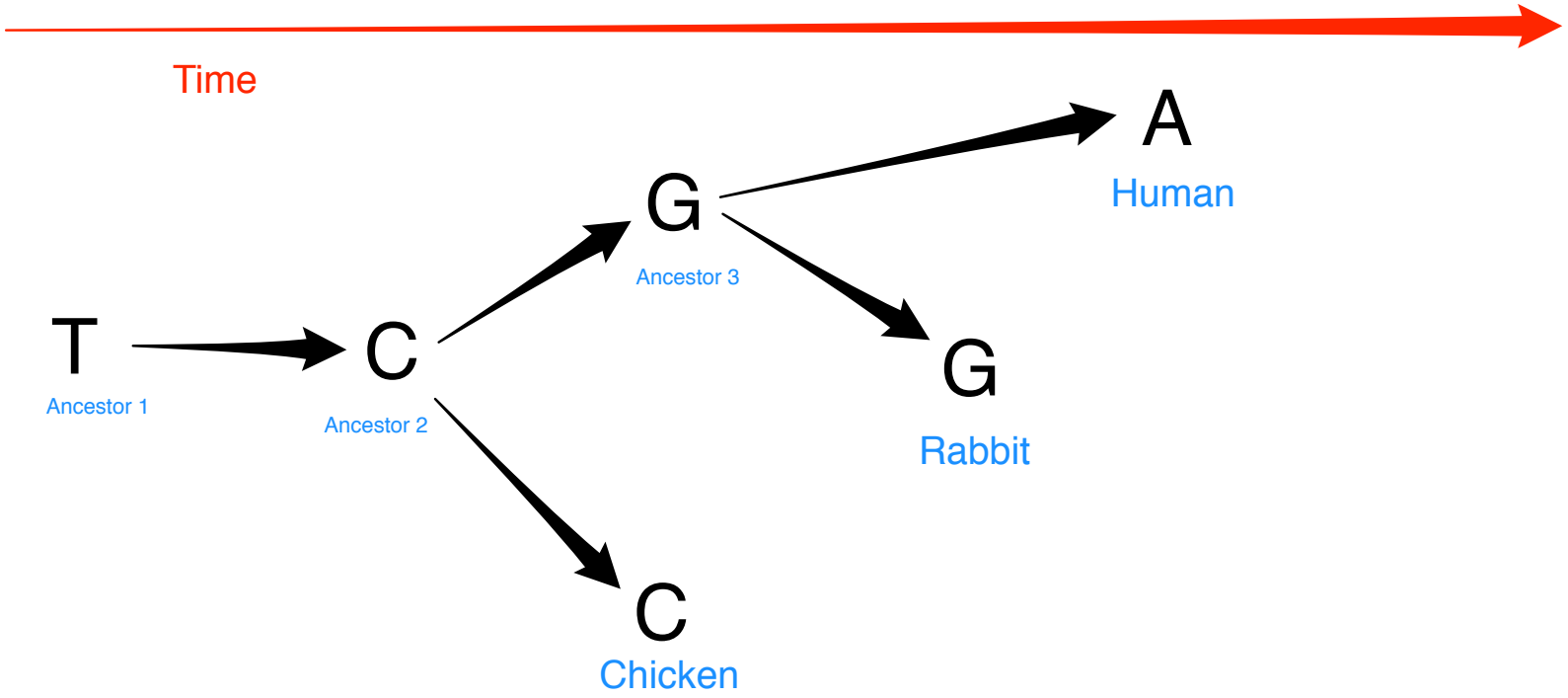
$$= \begin{bmatrix} D_1 & \cdots & D_C \end{bmatrix}$$



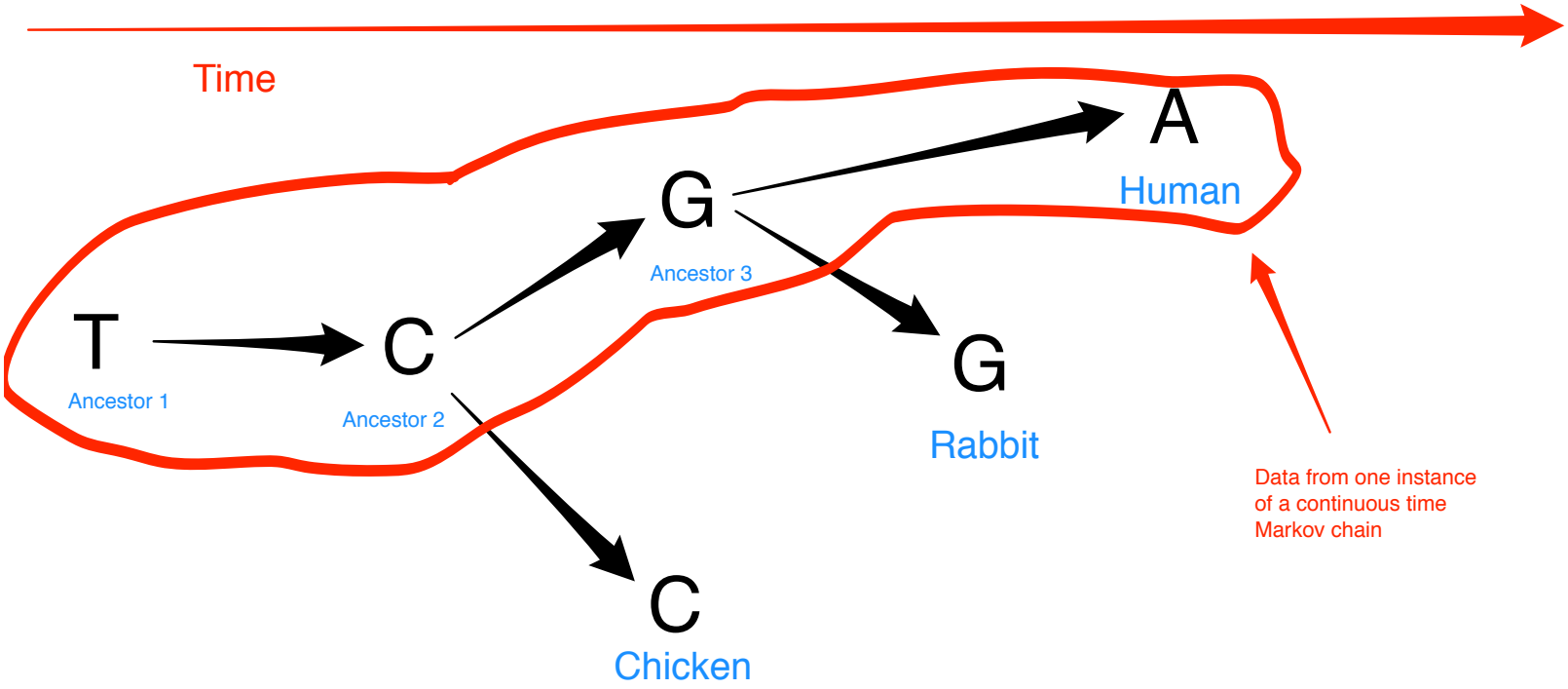

We assume that mutations in different loci are independent.

Consider locus 4...

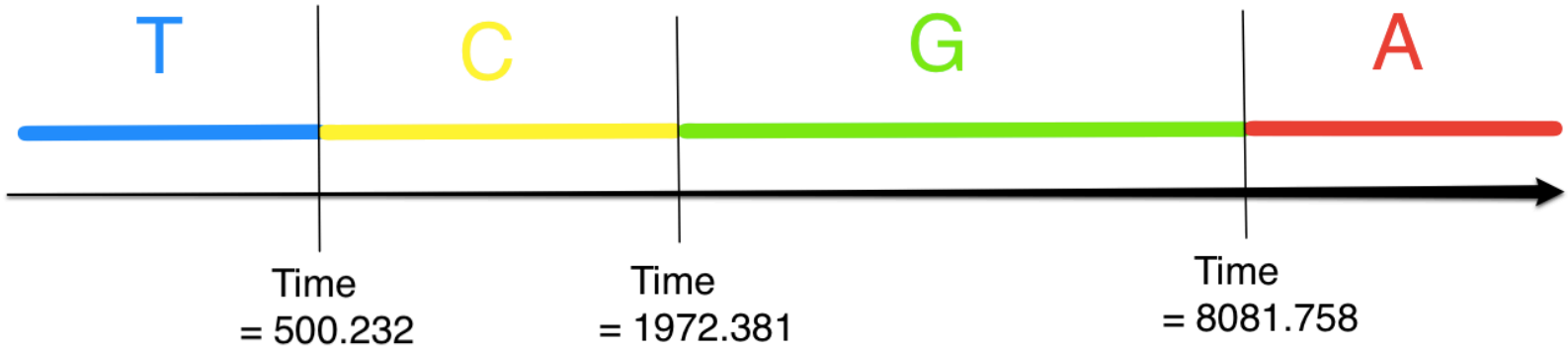
The nucleotide base in locus 4
might have changed over time in the following way:



We assume each path along the tree, or line of speciation, happens as a continuous time Markov chain:



We assume each path along the tree, or line of speciation, happens as a continuous time Markov chain:



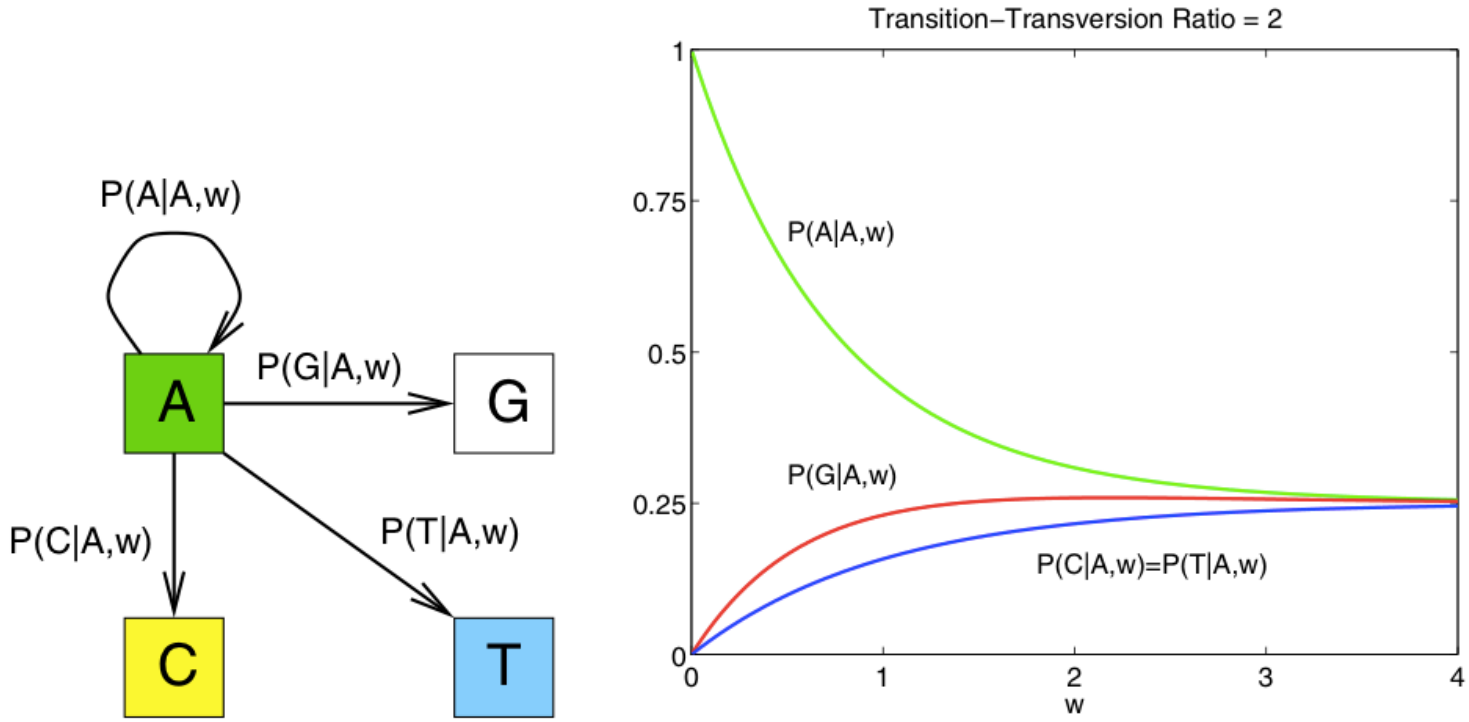



**Fig. 4.17. Mathematical model of nucleotide substitution.** *Left:* Nucleotide substitutions are modelled as transitions in a 4-element state space. The transition probabilities depend on the phylogenetic time $w = \lambda t$, where $t$ is physical time, and $\lambda$ is an unknown nucleotide substitution rate. *Right:* Dependence of the transition probabilities (vertical axis) on $w$ (horizontal axis). The graphs were obtained from the Kimura model (4.16) with a transition–transversion ratio of 2.

Each physical time length $t$ gives us a $4 \times 4$ transition matrix:

$$\mathbf{P}(t) = \begin{pmatrix} P(y(t) = A|y(0) = A) & \dots & P(y(t) = A|y(0) = T) \\ P(y(t) = G|y(0) = A) & \dots & P(y(t) = G|y(0) = T) \\ P(y(t) = C|y(0) = A) & \dots & P(y(t) = C|y(0) = T) \\ P(y(t) = T|y(0) = A) & \dots & P(y(t) = T|y(0) = T) \end{pmatrix}$$

Using calculus and some extra assumptions, we can describe $\mathbf{P}(t)$ by :

$$\mathbf{P}(t) = \exp(\mathbf{R}t)$$

where $\mathbf{R}$ is the **rate matrix**.

A possible design for $\mathbf{R}$, the so-called Kimura model:

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

where $\alpha$ and $\beta$ are parameters that make the columns sum to zero.

# DETERMINING THE BEST PHYLOGENETIC TREE

Given a set genetic sequence data $D$, use maximum likelihood to estimate the best:

1. CTMC model, $m$
2. phylogenetic tree, $\tau$

i.e., we want to find:

$$\widehat{\tau} = \text{argmax}_{\tau}\ P(D \mid \tau, m)$$

the MLE of the space of all possible phylogenetic trees.

The likelihood calculation is time-consuming using CPUs.

In their 2009 *Bioinformatics* paper, Suchard and Rambaut describe parallel, GPU-capable implementations of the following two steps in the workflow:

1. Given a tree, compute the probabilities of observing two specific sequences at either node of each edge.
2. Sum the data likelihood over all possible unobserved sequences at the internal nodes.

$$\Pr(\mathbf{s}, \mathbf{D} \mid \mathbf{r}) = \prod_{c=1}^{C} \Pr(\mathbf{s}_c, \mathbf{D}_c \mid r_c)$$

$$= \prod_{c=1}^{C} \left[ \pi_{s_{c1}} \prod_{b \in \mathcal{I}} P^{(r_c)}_{s_{c\psi(b)} s_{c\phi(b)}}(t_b) \right.$$

$$\left. \times \prod_{b \in \mathcal{E}} P^{(r_c)}_{s_{c\psi(b)} D_{c\phi(b)}}(t_b) \right]$$

$$\Pr(\mathbf{D}) = \prod_{c=1}^{C} \left[ \sum_{r=1}^{R} \left( \sum_{s_1=1}^{S} \cdots \sum_{s_{n-1}=1}^{S} \pi_{cs_1} \prod_{b \in \mathcal{I}} P^{(r)}_{s_{c\psi(b)} s_{c\phi(b)}}(t_b) \right. \right.$$
$$\left. \left. \times \prod_{b \in \mathcal{E}} P^{(r)}_{s_{c\psi(b)} D_{c\phi(b)}}(t_b) \right) \Pr(r) \right]$$

$$F_{urcs} = \left[\sum_{j=1}^{S} F_{\phi(b_1)rcj} \times P_{sj}^{(r)}(t_{b_1})\right] \times \left[\sum_{j=1}^{S} F_{\phi(b_2)rcj} \times P_{sj}^{(r)}(t_{b_2})\right]$$

$$\Pr(\mathbf{D}) = \prod_{c=1}^{C} \left[ \sum_{r=1}^{R} \left( \sum_{s_1=1}^{S} \pi_{s_1} \times F_{1rcs_1} \right) \Pr(r) \right]$$

$$\begin{aligned}\mathbf{P}^{(r)}(t) &= \exp(\mu_r t\mathbf{\Lambda}) = \mathbf{E} \times \text{diag}(e^{\mu_r t\lambda_1}, \ldots, e^{\mu_r t\lambda_S}) \times \mathbf{E}^{-1} \\ &= \mathbf{E}\mathbf{D}_{rt}\mathbf{E}^{-1}\end{aligned}$$