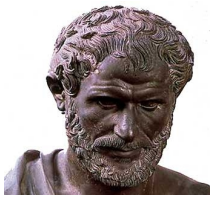
O desenvolvimento da Cladística

Dados moleculares e modelos probabilísticos

Aristóteles – 384-322 A.C.



Darwin
1809-1882

1859

Período essencialista

Mundo dinâmico

Resistência e Nova Síntese

Sistemática Evolutiva

1936 - 1947

1960's

Fenética

1970's

Cladística

1990's

Probabilismo

Carolus Linnaeus
1707-1778

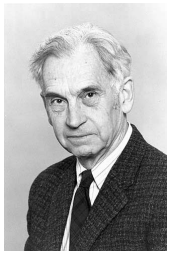
Buffon
1707-1788

Lamarck
1744 -1829

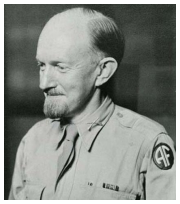
St-Hilair
1772 -1844

Cuvier
1769 -1832

Ernest Mayr
1904 - 2005

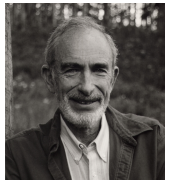


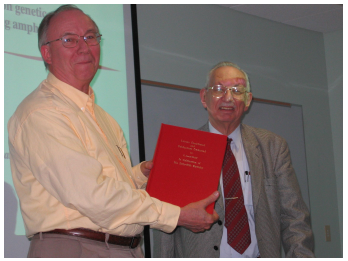
G.G. Simpson
1902 - 1984



Theodosius Dobzhansky
1900 -1975

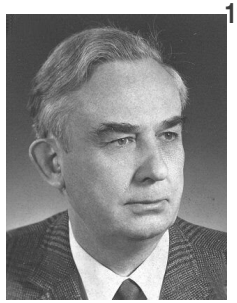
Paul Erlich





James Rohlf

R. Sokal
1926 -



Willi Hennig
1913 - 1976

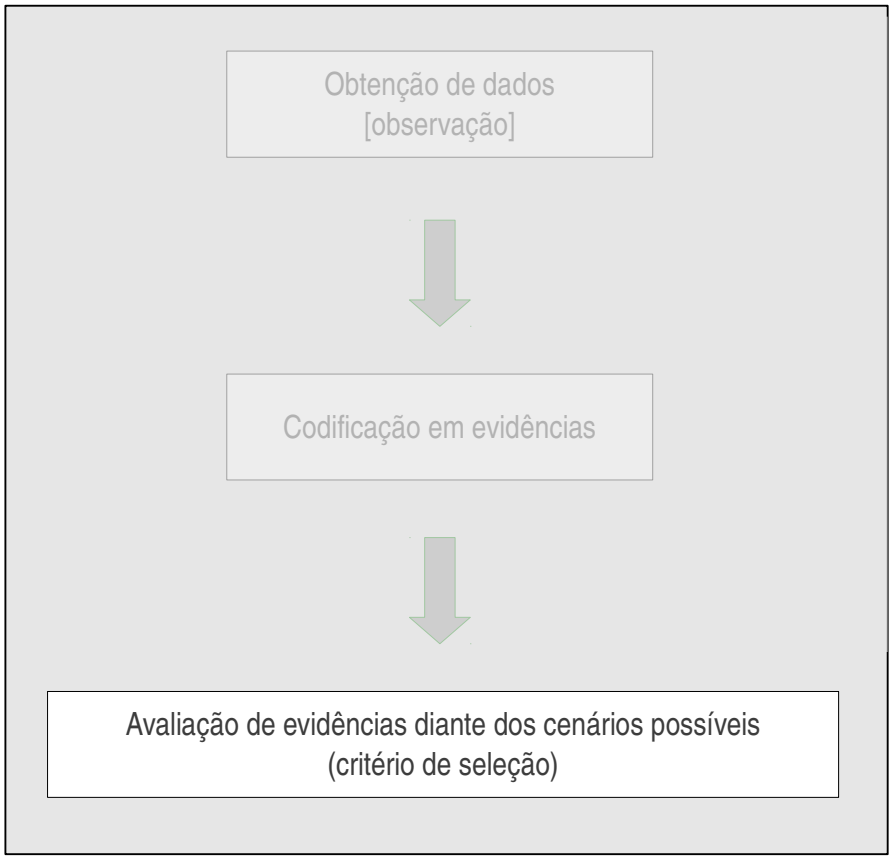
Steve Farris

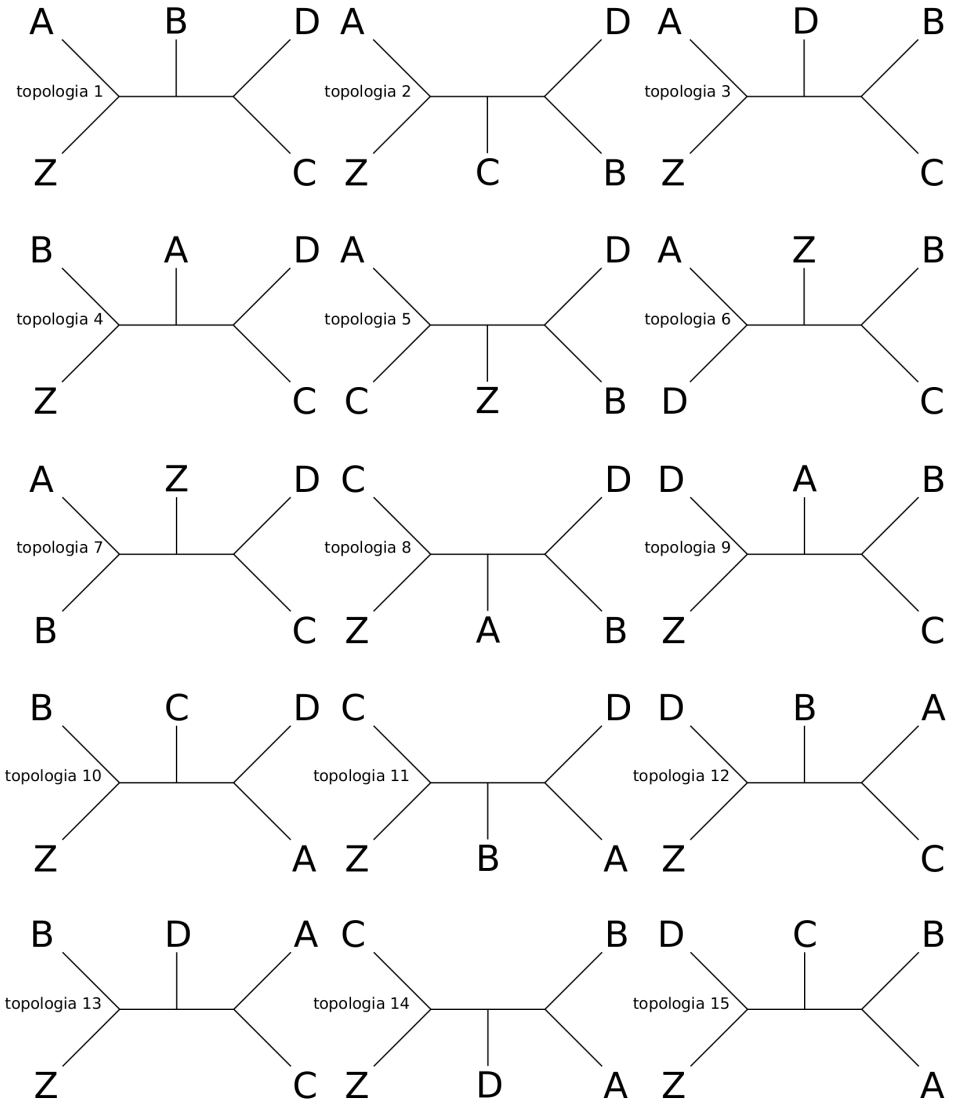
Joe Felsenstein

David Hillis

Lógica da inferência filogenética

Avaliação e critério de seleção: soluções possíveis





Qual diagrama explica melhor minhas observações, ou que considero como evidência?

Lógica da inferência filogenética

Avaliação e critério de seleção: soluções possíveis

Obtenção de dados
[observação]

Codificação em evidências

Avaliação de evidências diante dos cenários possíveis
(critério de seleção)

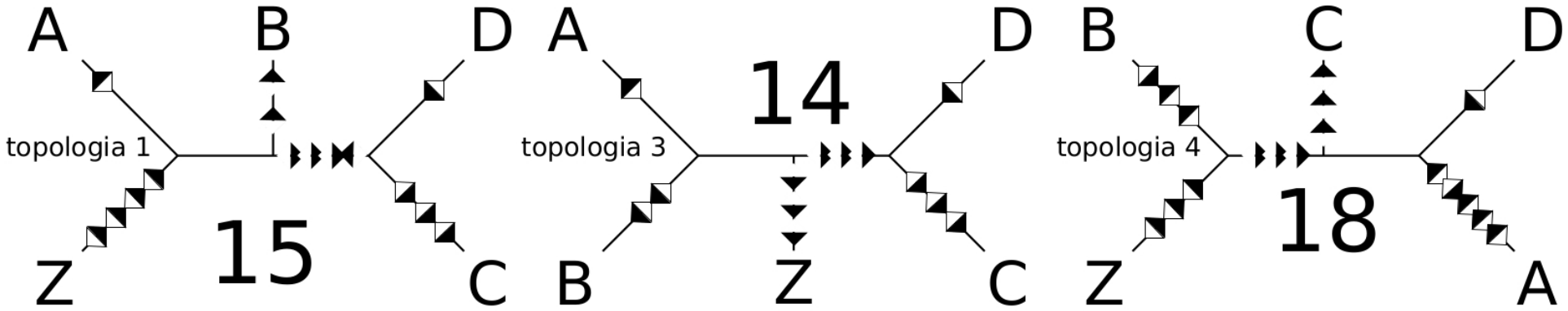
Critério de seleção: **parcimônia**

William of Ockham (c. 1288 - c. 1348):
lex parsimoniae ou “Occam's Razor”

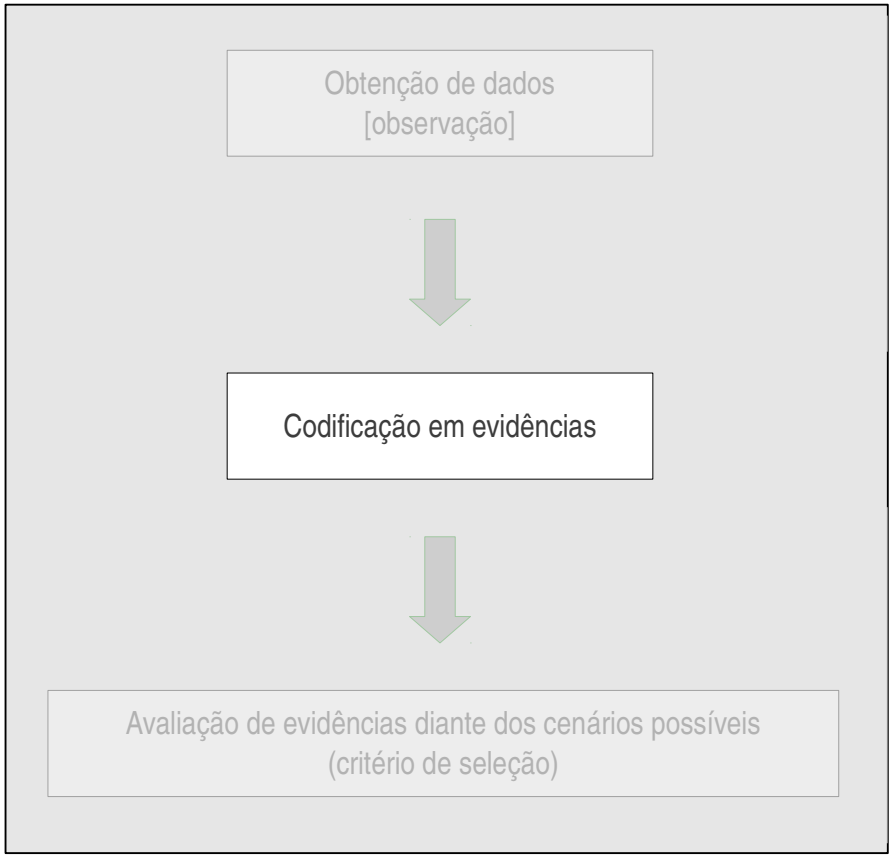
"entia non sunt multiplicanda praeter necessitatem"

"Entities should not be multiplied unnecessarily."

"when you have two competing theories which make exactly the same predictions, the one that is simpler is the better."



Lógica da inferência filogenética




```
sp.X   CTGGCTACGT
sp.A   TGGAGTAAGT
sp.B   CCTAGCAAGT
sp.C   CCTGATTGCA
```

Parcimônia:

EVIDÊNCIAS: transformação de estados de caracteres

```
 1a. posição: C ↔ T
 3a. posição: G ↔ T
10a. posição: T ↔ A
```

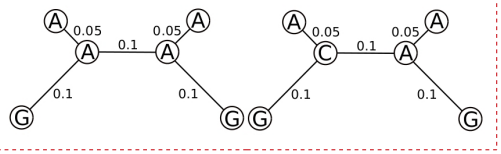
CODIFICAÇÃO: matriz de dados

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
sp.X	1	3	2	2	1	3	0	1	2	3
sp.A	3	2	2	0	2	3	0	0	2	3
sp.B	1	1	3	0	2	1	0	0	2	3
sp.C	1	1	3	2	0	3	3	2	1	0

Probabilística (ML):

EVIDÊNCIAS: modelo de transformações + topologia que melhor explicam seus dados.

$$\begin{bmatrix} P_{(AA)} & P_{(AC)} & P_{(AG)} & P_{(AT)} \\ P_{(CA)} & P_{(CC)} & P_{(CG)} & P_{(CT)} \\ P_{(GA)} & P_{(GC)} & P_{(GG)} & P_{(GT)} \\ P_{(TA)} & P_{(TC)} & P_{(TG)} & P_{(TT)} \end{bmatrix} +$$

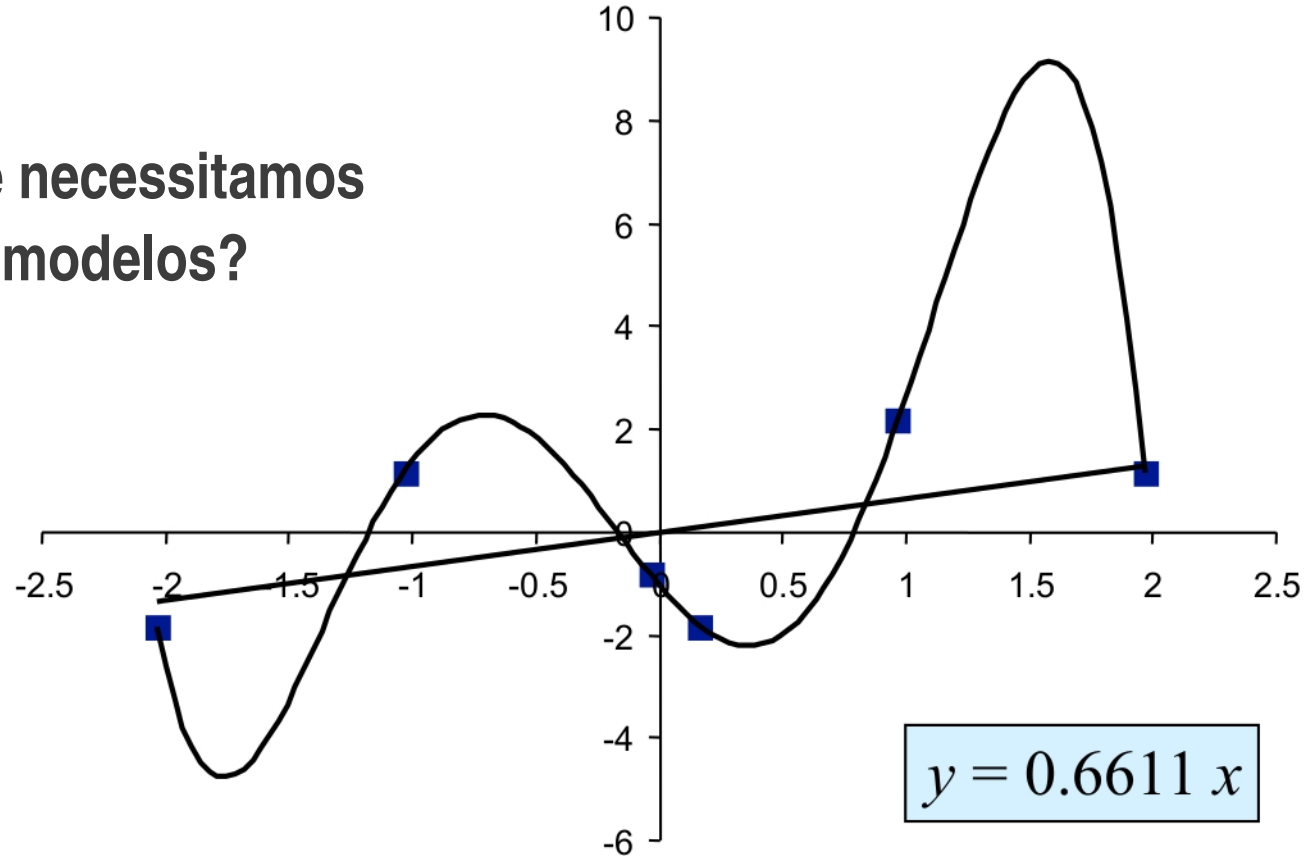



```
sp.X CTGGCTACGT
sp.A TGGAGTAAGT
sp.B CCTAGCAAGT
sp.C CCTGATTGCA
```

Modelos

$$y = -1.5972\,x^5 + 23.167\,x^4 - 126.18\,x^3 + 319.17\,x^2 - 369.22\,x + 155.67$$

Por que necessitamos de modelos?



Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

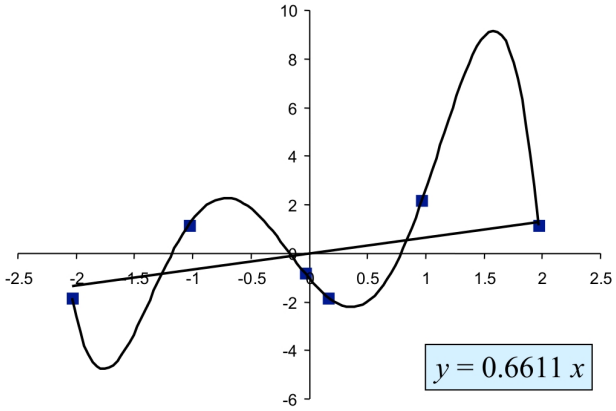
Modelos

Models and Representation: Models can perform two fundamentally different representational functions. On the one hand, a model can be a representation of a selected part of the world (the 'target system'). Depending on the nature of the target, such models are either models of phenomena or models of data. On the other hand, a model can represent a theory in the sense that it interprets the laws and axioms of that theory. These two notions are not mutually exclusive as scientific models can be representations in both senses at the same time.

Modelo probabilistico: "*is an explicit model of potential observations that includes a description of the uncertainty of those observations due to natural variation, to errors in measurements, or to complete information, [...].*

"In fact, scientists often find that complex models do very poorly in predicting new data when fitted to old. Simpler models often do better. Here the complexity of a model corresponds to the number of adjustable parameters it contains. "(Sober, 2008:82)

$y = -1.5972\,x^5 + 23.167\,x^4 - 126.18\,x^3 + 319.17\,x^2 - 369.22\,x + 155.67$



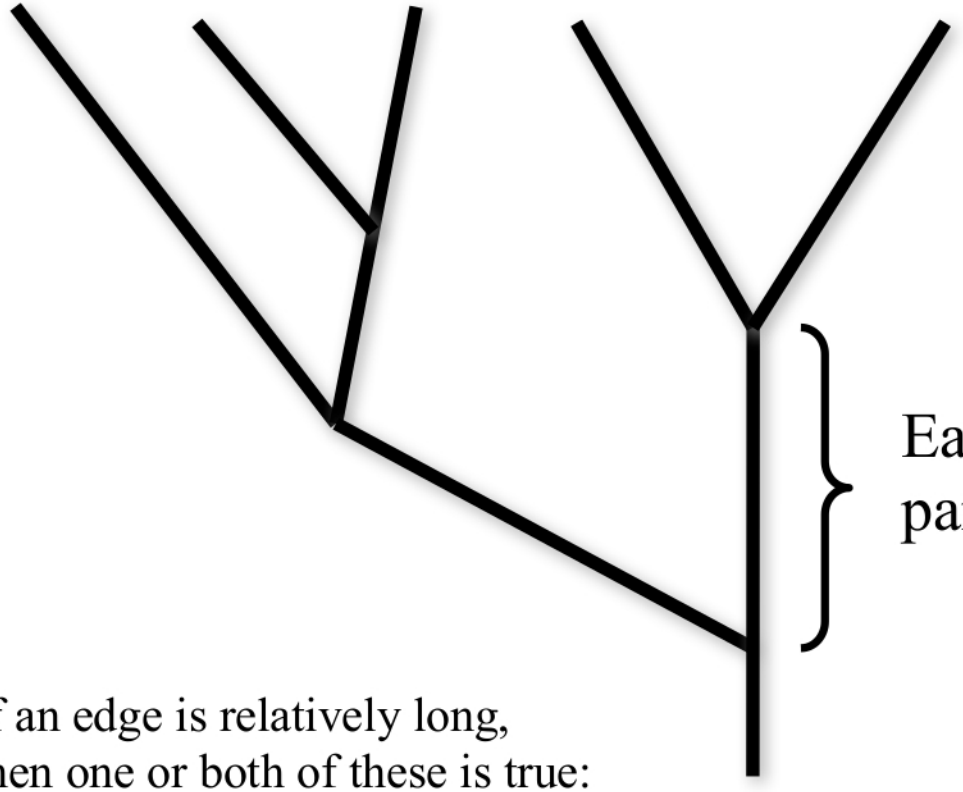
Modelos, Realismo e Instrumentalismo

"The problem of evaluating how accurately models predict new data when fitted to old has a philosophically interesting property: a model known to be false will sometimes be more predictively accurate than a model known to be true. […]

Instrumentalism is the view that the goal of scientific inference is to find theories that make accurate predictions, not to find theories that are true. It stands opposed to **scientific realism**, which holds that the goal is to find true theories. […]

The philosophical debate concerns what **scientific inference** is able to attain, not what scientists yearn for. If the inference procedures used in science are able to discover which theories are true, or which are probably true, then realism is cor- rect. If those procedures are capable only of discovering which theories will make the most accurate predictions, then instrumentalism is." (Sober, 2008:97)

Modelos de substituição



An edge length represents the *expected number of substitutions*, which equals the overall *substitution rate* (3α) multiplied by *time* (t)

$$\nu = (3\alpha)t$$

(Jukes-Cantor model)

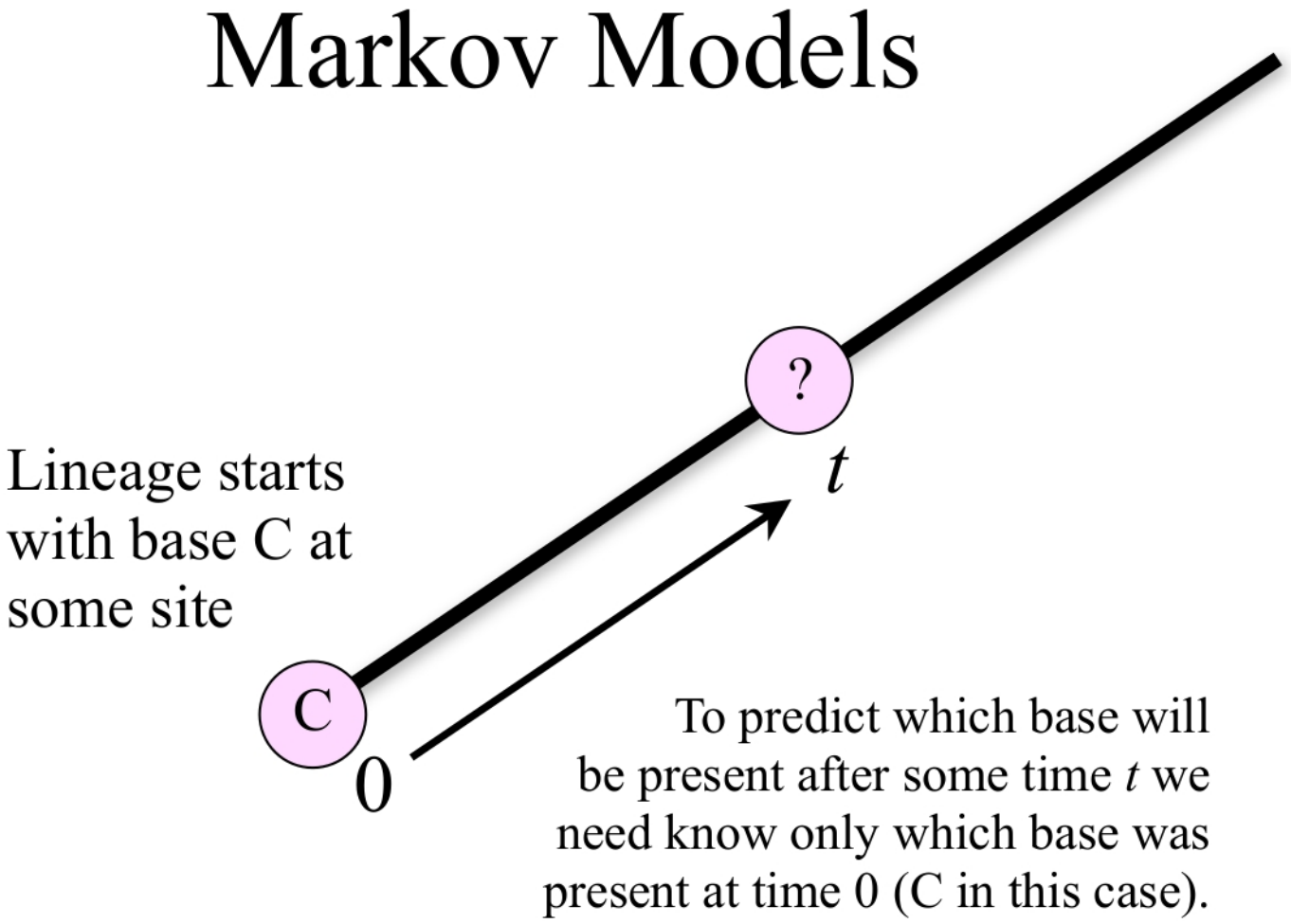
Each edge length ν is a parameter in the model

If an edge is relatively long, then one or both of these is true:

- the substitution rate was high
- the lineage was in existence for a long time

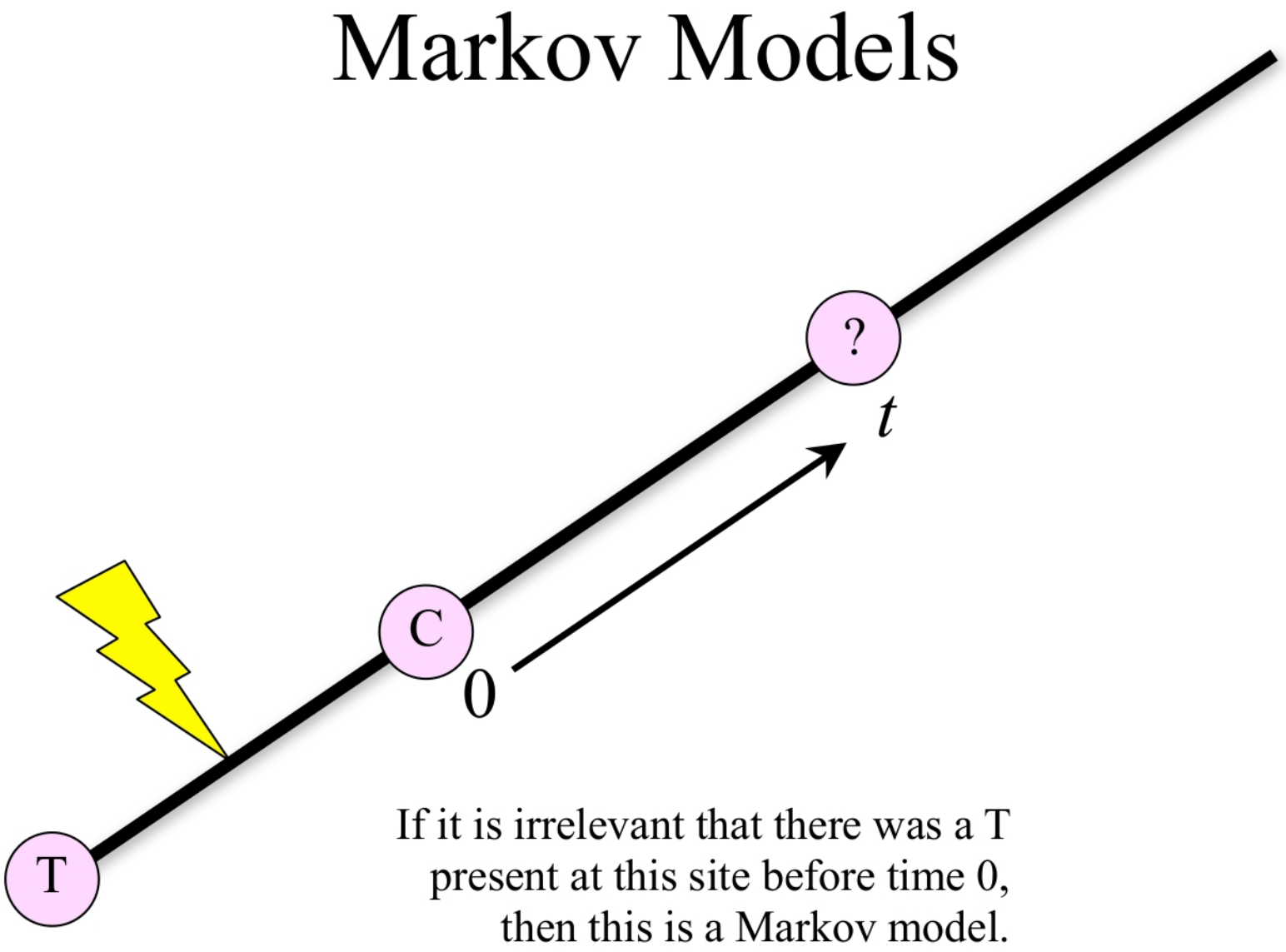
Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Modelos de substituição



Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

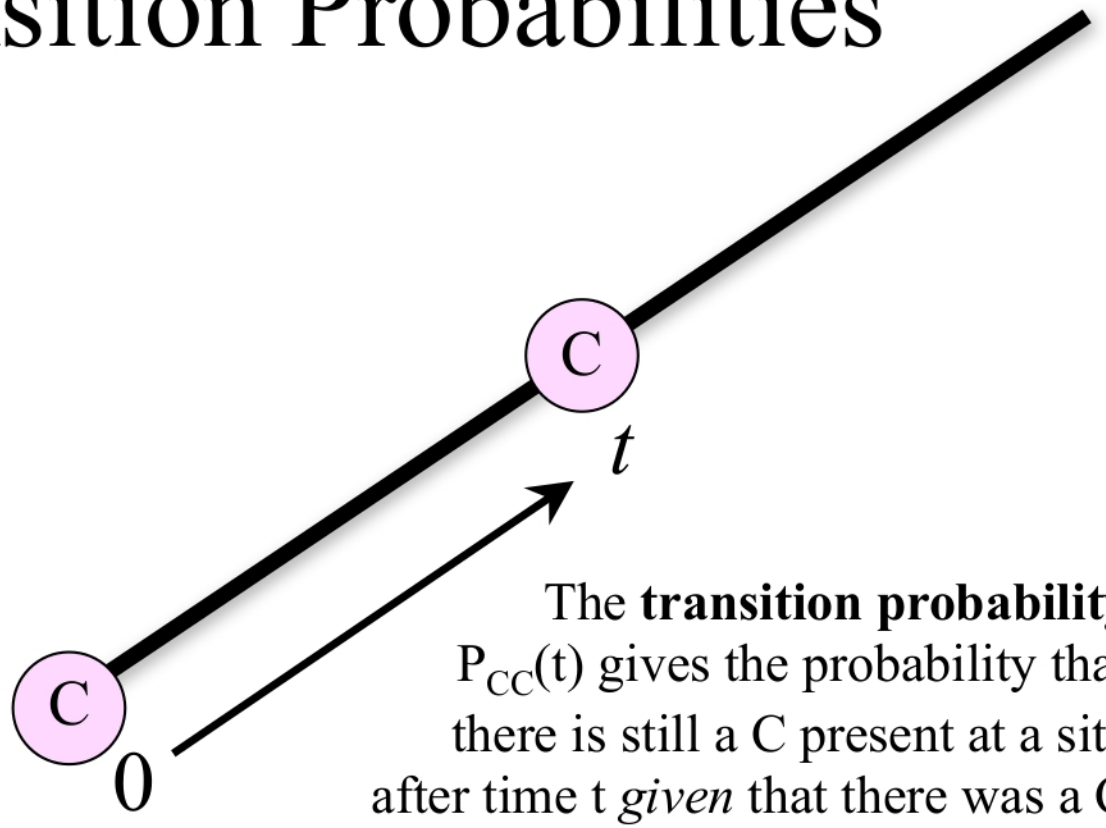
Modelos de substituição



Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Modelos de substituição

Transition Probabilities



The **transition probability** $P_{CC}(t)$ gives the probability that there is still a C present at a site after time t *given* that there was a C present at time 0

Note: the term *transition* here comes from the terminology of stochastic processes and refers to any change of state (and even non-changes!). This kind of transition could thus be a transition-type or a transversion-type substitution if the states were taken to be the four nucleotides

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

JC Transition Probability

Here is the probability that a site starting in state T will end up in state G after time t when the substitution rate is α:

$$P_{TG}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$

(the symbol e is the base of the natural logarithms and is thus a constant: 2.718281828459045...)

There is only one unknown parameter here: the edge length ν. Since $\nu = 3\alpha t$, let's work with the following identical formula:

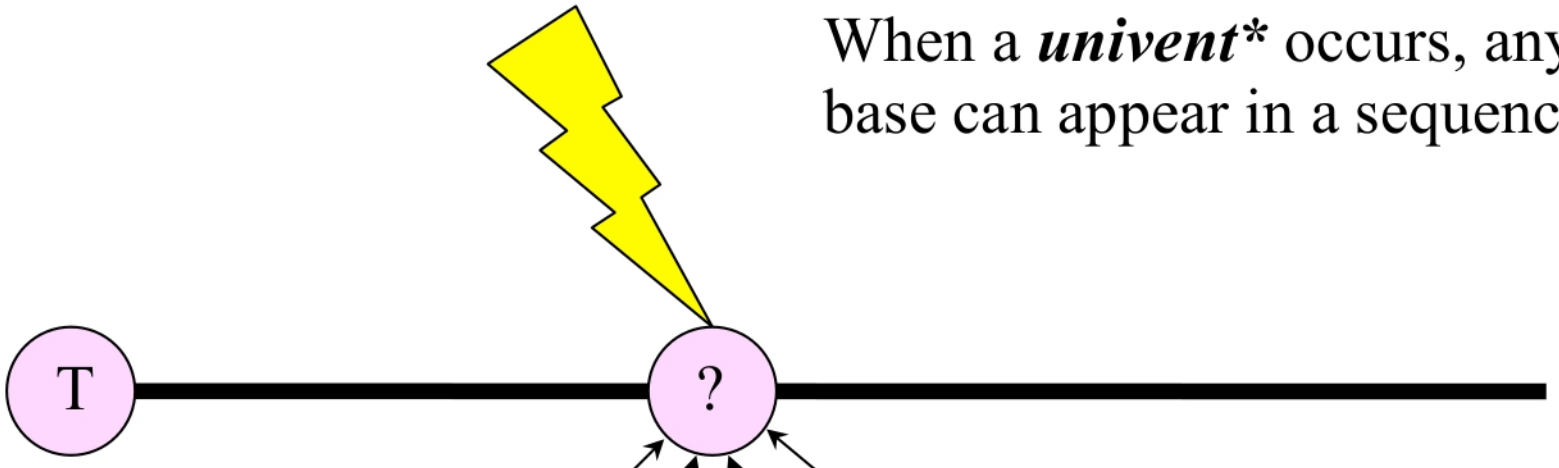
$$P_{TG}(t) = \frac{1}{4}\left(1 - e^{-4\alpha t}\right)$$

Where does a transition probability formula such as this come from?

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Modelos de substituição

"Univents" vs. substitutions



When a ***univent**** occurs, any base can appear in a sequence.

If the base that appears is *different* from the base that was already there, then a **substitution** event has occurred.

The rate (α) at which any *particular* substitution occurs will be 1/4 the univent rate (μ). That is, $\alpha = \mu/4$

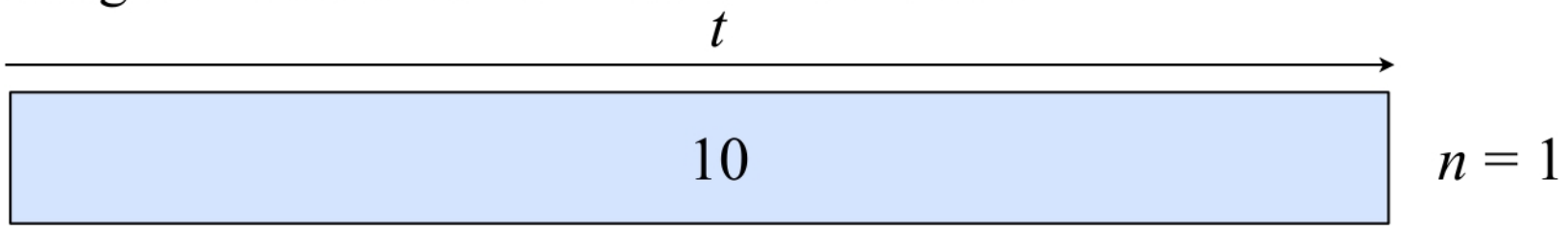
*You will not find the word *univent* in the literature; however, this concept plays an important role in a technique called *uniformization*, which hold some promise for making complex models more practical.

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Modelos de substituição

Poisson probabilities

Over a branch that has been in existence for some time period t, imagine that there have been 10 univents...



Now divide the time interval by 2...

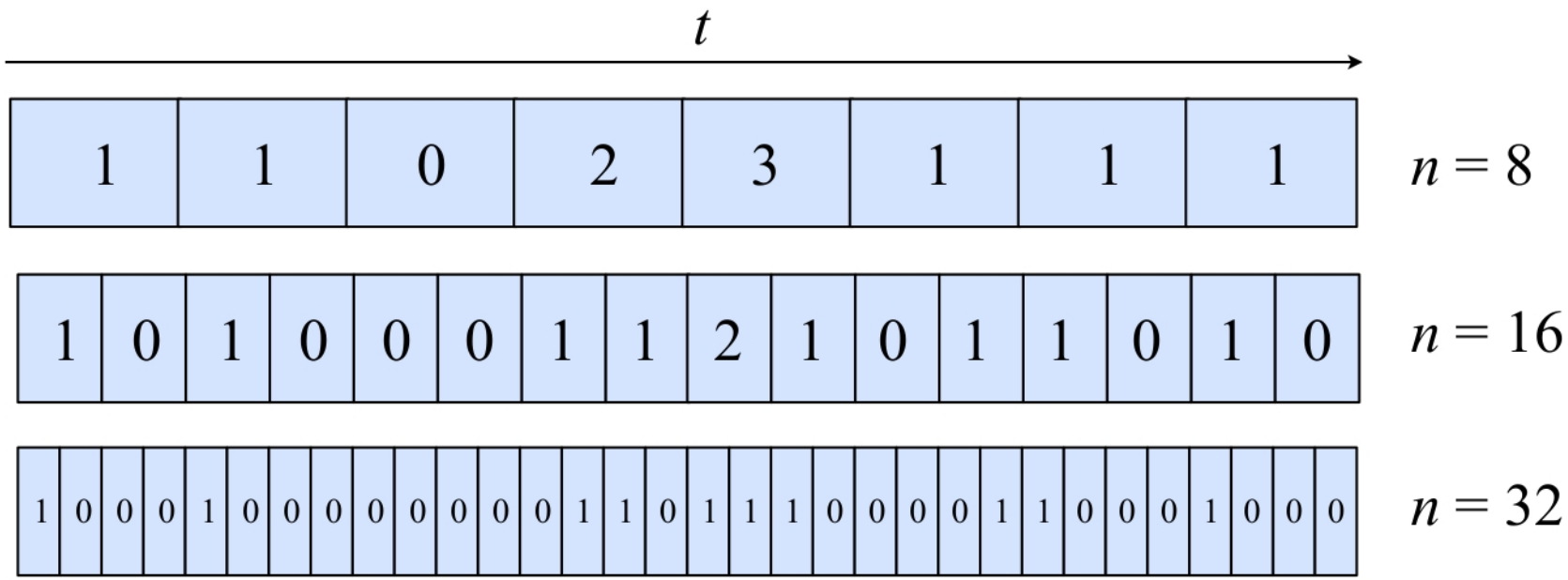
Then divide it by two again...

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Modelos de substituição

Poisson probabilities

Keep on dividing it until no interval has more than 1 substitution...



Now there are only two categories of intervals:

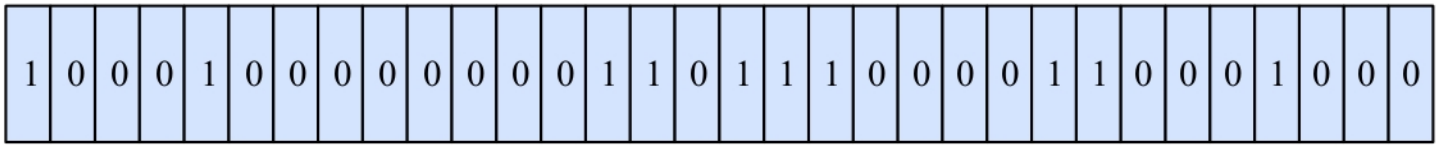
y = 10 intervals contain 1 univent

n - y = 22 intervals contain 0 univents

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Modelos de substituição

Poisson probabilities



$n = 32$

We can treat this as a series of n = 32 independent trials, each of which resulted in either a success (1) or a failure.

If we let the probability of success on each trial be *p*, the probability of failure is 1 - *p* and the probability of 10 univents is given by the binomial probability formula:

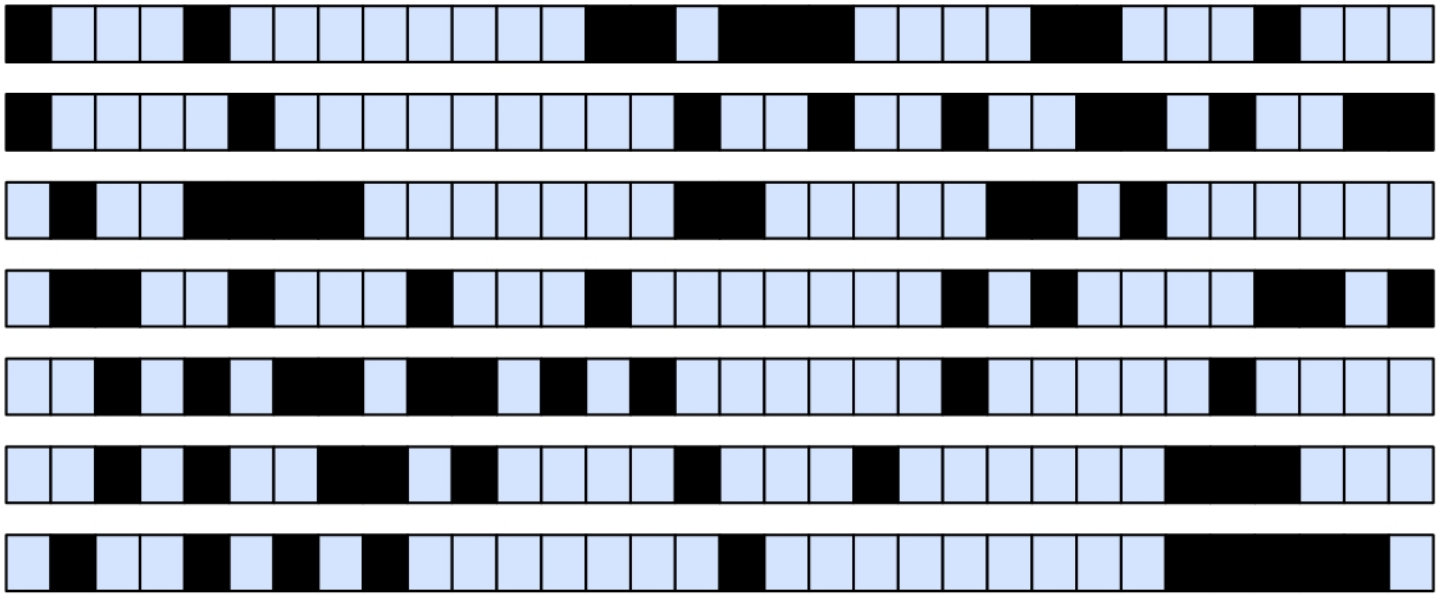
$$\Pr(y|p) = \binom{n}{y} p^y (1-p)^{n-y}$$

where the first term is the number of ways of rearranging the y 1s amongst the n bins (see next slide).

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Modelos de substituição

Poisson probabilities



The formula below provides the number of different ways that y=10 univents could be distributed amongst n=32 bins.

$$\binom{n}{y} = \frac{n!}{y!\,(n-y)!} = \frac{32!}{10!\,22!} = 64512240$$

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Modelos de substituição

Poisson probabilities

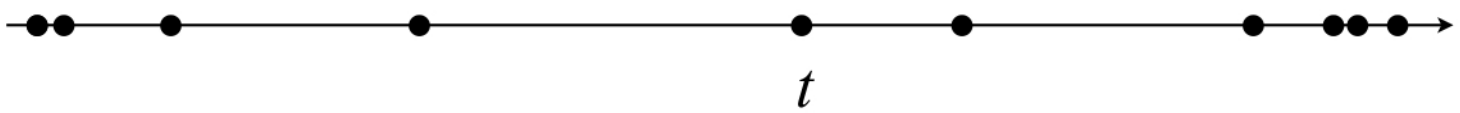
Now imagine continuing to **double** the number of bins while **halving** p so that the expected total number of univents (λ) stays constant. This amounts to keeping the product of n and p constant:

$$\lambda = np = (2n)\left(\frac{p}{2}\right)$$

Replacing p with the equivalent value λ/n and finding the limit as n is increased without limit yields

$$\Pr(y|\lambda) = \lim_{n\to\infty} \left\{ \left(\frac{n!}{y!\,(n-y)!} \right) \left(\frac{\lambda}{n}\right)^{y} \left(1 - \frac{\lambda}{n}\right)^{n-y} \right\}$$

We have now divided our time t into infinitely many bins, and the intervals in which a univent has occurred appear as points along a continuous time line:



Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Poisson probabilities

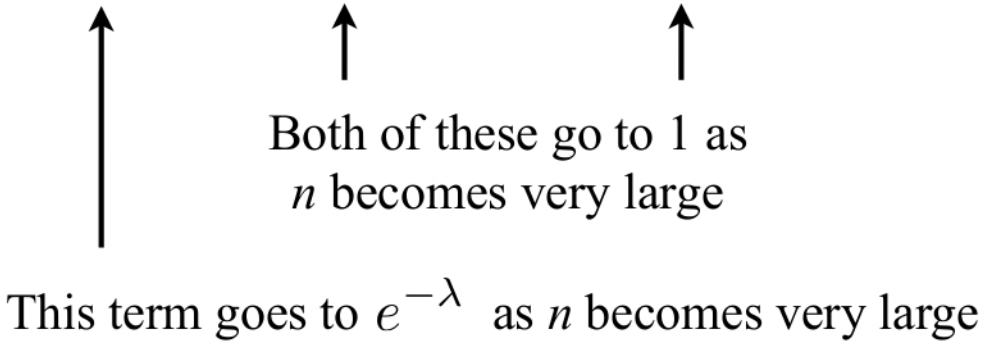
$$\Pr(y|\lambda) = \lim_{n\to\infty} \left(\frac{n!}{y!\ (n-y)!}\right) \left(\frac{\lambda}{n}\right)^{y} \left(1-\frac{\lambda}{n}\right)^{n-y}$$

Poisson probability formula:

$$\Pr(y|\lambda) = \frac{\lambda^y}{y!} e^{-\lambda}$$

Important special cases:

$$\Pr(y=0|\lambda) = e^{-\lambda}$$
$$\Pr(y>0|\lambda) = 1-e^{-\lambda}$$



The quantity λ is the expected number of univents across the interval of length t. We can thus replace λ by the univent rate μ times the time t: $\lambda = \mu\, t$

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Deriving a transition probability

Calculate the probability that a site currently T will change to G over time *t* when the rate of this particular substitution is α:

Pr(zero univents) = $e^{-\mu t}$

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

JC69 model

- Bases are assumed to be equally frequent (all 0.25)
- Assumes rate of substitution (α) is the same for all possible substitutions
- Usually described as a 1-parameter model (the parameter being ν)
- Remember, however, that each edge in a tree can have its own ν, so there are really as many parameters in the model as there are edges in the tree!

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Transition Probabilities: Remarks

$$P_{TA}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$

$$P_{TC}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$

$$P_{TG}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$

$$P_{TT}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$

$$= 1 - e^{-4\nu/3}$$

Oops! Should be 1.0 because T must either stay the same or change to A, C or G. What have I forgotten?

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Transition Probabilities: Remarks

$$P_{TA}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$

$$P_{TC}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$

$$P_{TG}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$

$$P_{TT}(t) = e^{-4\nu/3} + \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$

$$= 1$$

Forgot to account for the possibility that the nucleotide could *stay the same* if there were *zero* univents over time *t*

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

More on Transition Probabilities

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\nu/3}$$

Different state at both ends

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\nu/3}$$

Same state at both ends

What are the transition probabilities if $\nu = \infty$?

$$P_{ij}(\infty) = P_{ii}(\infty) = \frac{1}{4}$$

What are the transition probabilities if $\nu = 0$?

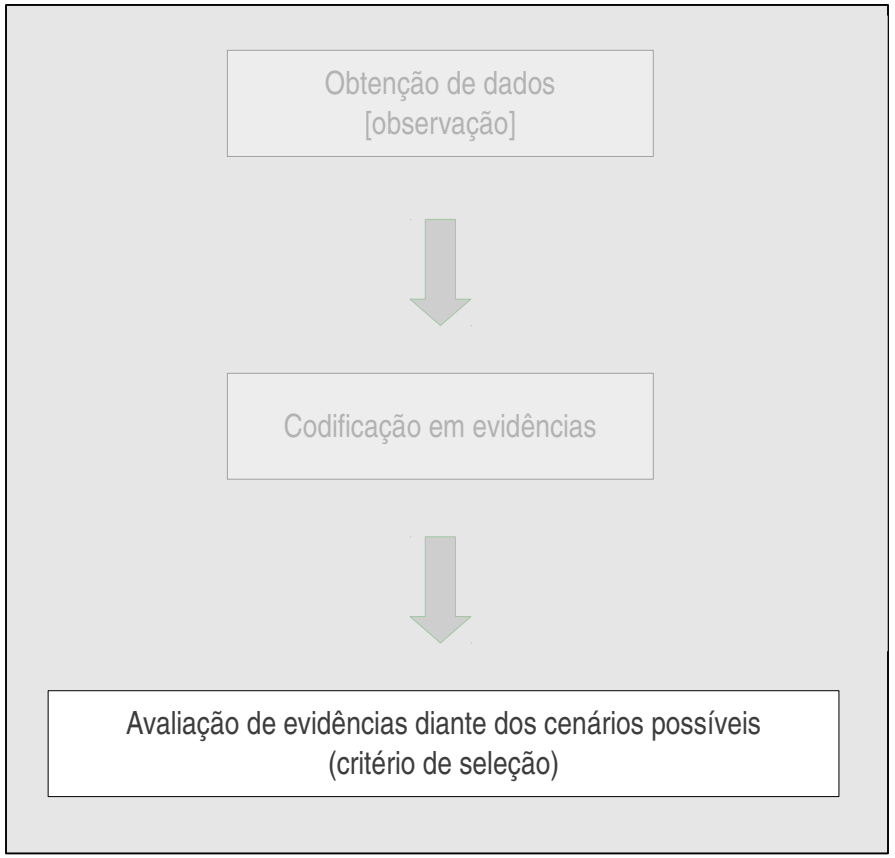
$$P_{ij}(0) = 0 \quad P_{ii}(0) = 1$$

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Lógica da inferência filogenética

Avaliação e critério de seleção: soluções possíveis

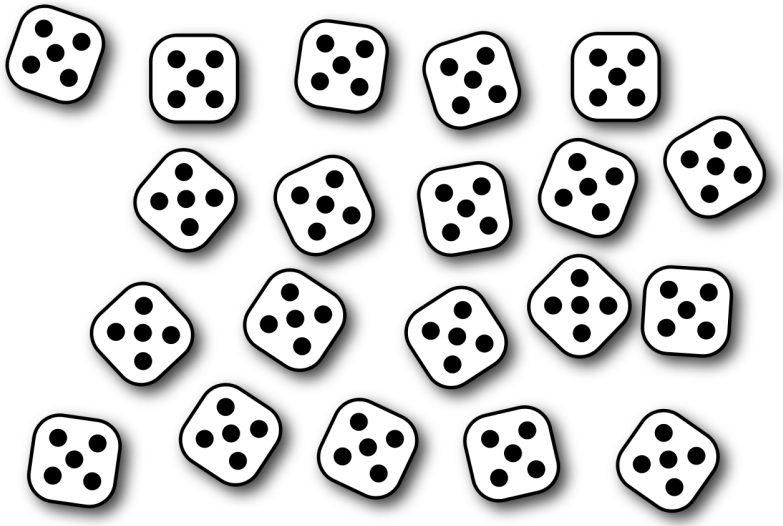
Critério de seleção: **verossimilhança (*Likelihood*)**



A probabilidade das observações calculadas utilizando **um modelo** nos diz o quão surpresos nós estaríamos com os dados observados

O modelo escolhido é aquele que menos surpreende!

Considere que eu jogue 20 dados sobre uma mesa e obtenha o seguinte resultados:



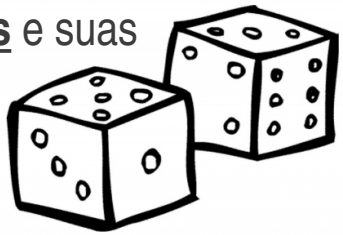
Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Probabilidades e modelos

Avaliação e critério de seleção: soluções possíveis

Modelo probabilistico: Representação matemática de um fenômeno aleatório.

Se dois eventos não possuem soluções em comum, eles são chamados de **disjuntos** e suas probabilidades obedecem à seguinte regra:



Regra 3: Se dois eventos A e B são disjuntos, então a probabilidade de qualquer um destes eventos e a soma das probabilidades dos dois eventos: $P_{(A\ ou\ B)} = P_{(a)} + P_{(b)}$

Regra 4: A probabilidade de um evento ***A*** não ocorrer é $P_{(A^c)} = 1 - P_{(A)}$

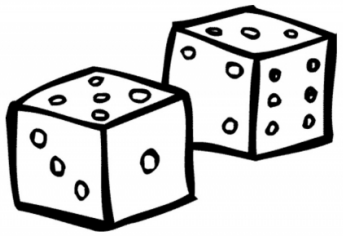
$P_{(\boxdot)} = A/S = 1/6$

$P_{(\boxdot^c)} = 1 - P_{(A)} = 1 - 1/6$

Fonte: http://www.stat.yale.edu/Courses/1997-98/101/probint.htm

Probabilidades e modelos

Considere que você jogue dois dados:



A probabilidade de obter “olhos de serpente” é expressa pela seguinte regra:

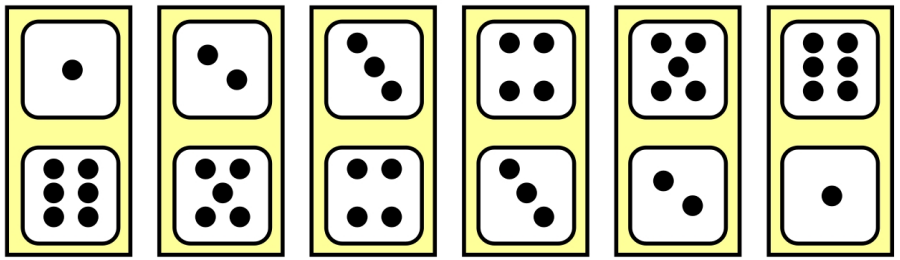
Regra 5: Se dois eventos A e B são independentes, então a probabilidade dos dois eventos ocorrerem um após o outro (i.e., intersecção de eventos, ou probabilidade conjunta) é o produto das probabilidades de cada evento: $P_{(A\ e\ B)} = P_{(A)} * P_{(B)}$.

Combinando as regras:

Qual a probabilidade de se obter a soma 7 ao jogar dois dados?

E

1/6 * 1/6



(1/36)+(1/36)+(1/36)+(1/36)+(1/36)+(1/36) = 1/6

OU OU OU OU OU

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Verossimilhança e Modelos

$$L_{(\text{mod.} \mid \text{obs.})} = \Pr(\text{observação} \mid \text{modelo})$$

Modelo probabilistico: "*is an explicit model of potential observations that includes a description of the uncertainty of those observations due to natural variation, to errors in measurements, or to complete information, [...].*"

Fonte:Lewin-Koh, Taper & Lele (2004 - A brief tour of statistical concepts. In: The nature of scientific evidence (eds. ML Taper and SR Lele), University of Chicago Press, pp 3 -16.)

Verossimilhança e Probabilidade

1. Suponha que você esteja ouvindo um barrulho no forro de sua casa.

2. Você considera a hipótese de que há *gremlins* jogando boliche no seu forro.

A verossimilhança desta hipótese é muito alta, pois se há *gremlins* jogando boliche no seu forro, há a probabilidade de have barrulho. No entanto, certamente você não pensa que o barrulho é evidência (torna-se provável) que haja *gremlins* jogando boliche no seu forro.

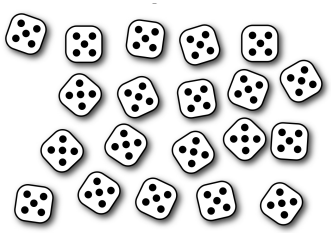
$$P_{(O|H)} = \text{muito alta}$$

$$P_{(H|O)} = \text{muito baixa}$$

Fonte: E. Sober. 2008. Evidence and Evolution: the logic behind the science. Cambridge Unibersity Press. 414 pp.

Verossimilhança e Modelos

"*Fair dice model*"

$$\Pr\left(\text{  } \mid \text{modelo Honesto}\right) = (1/6)^{20} =$$

$$\frac{1}{3.656.158.440.062.976}$$

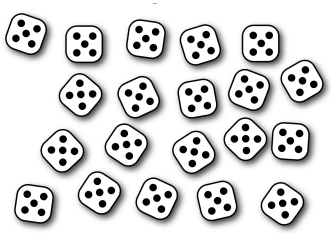
*Deveríamos estar muito **surpresos** com esta observação uma vez que as chances, ou probabilidade, deste evento é muito pequena: 1 em 3,6 quatrilhões!*

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Verossimilhança e Modelos

"'**Carlinhos Cachoeira**' *dice model*"

(Assume que todos os dados apresentam 5 em todos os lados)

$$\Pr\left(\text{  } \mid \text{modelo de C.C.} \right) = 1^{20} = 1$$

*Não deveríamos estar muito **surpresos** com esta observação uma vez que as chances, ou probabilidade, deste evento é muito grande: 1*

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

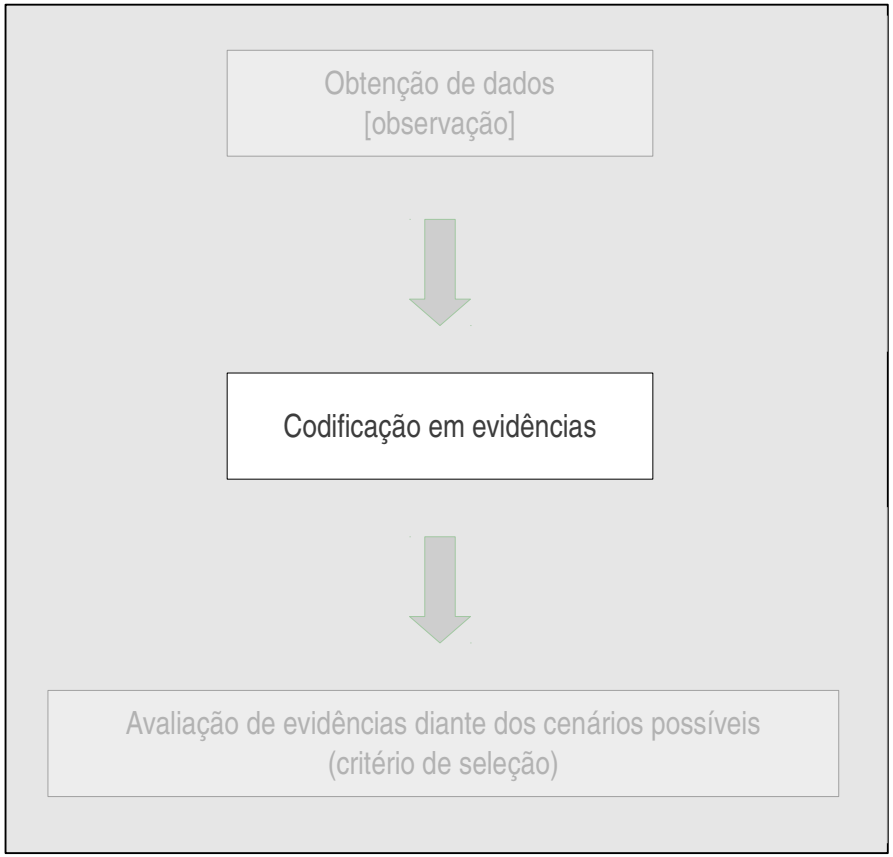
Verossimilhança Máxima: seleção

Modelo	Likelihood	Nível de surpreendimento
"*Fair dice model*"	1/3.656.158.440.062.976	muito, mas muito surpreso
"*Carlinhos Cachoeira dice model*"	1.0	nem um pouco surpreso

Modelo escolhido, pois maximiza a verossimilhança (*minimiza a surpresa*)

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Lógica da inferência filogenética




```
sp.X   CTGGCTACGT
sp.A   TGGAGTAAGT
sp.B   CCTAGCAAGT
sp.C   CCTGATTGCA
```

Parcimônia:

EVIDÊNCIAS: transformação de estados de caracteres

```
 1a. posição: C ↔ T
 3a. posição: G ↔ T
10a. posição: T ↔ A
```

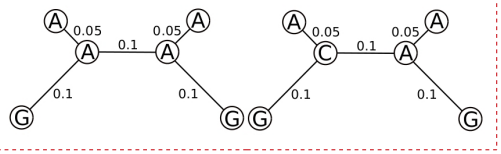
CODIFICAÇÃO: matriz de dados

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
sp.X	1	3	2	2	1	3	0	1	2	3
sp.A	3	2	2	0	2	3	0	0	2	3
sp.B	1	1	3	0	2	1	0	0	2	3
sp.C	1	1	3	2	0	3	3	2	1	0

Probabilística (ML):

EVIDÊNCIAS: modelo de transformações + topologia que melhor explicam seus dados.

$$\begin{bmatrix} P_{(AA)} & P_{(AC)} & P_{(AG)} & P_{(AT)} \\ P_{(CA)} & P_{(CC)} & P_{(CG)} & P_{(CT)} \\ P_{(GA)} & P_{(GC)} & P_{(GG)} & P_{(GT)} \\ P_{(TA)} & P_{(TC)} & P_{(TG)} & P_{(TT)} \end{bmatrix} +$$




```
sp.X CTGGCTACGT
sp.A TGGAGTAAGT
sp.B CCTAGCAAGT
sp.C CCTGATTGCA
```

Lógica da inferência filogenética: Likelihood

$$L = \Pr\left(\begin{matrix}\text{sp.X } CTGGCTACGT\\ \text{sp.A } TGGAGTAAGT\\ \text{sp.B } CCTAGCAAGT\\ \text{sp.C } CCTGATTGCA\end{matrix} \;\middle|\; \begin{bmatrix} P_{(AA)} & P_{(AC)} & P_{(AG)} & P_{(AT)}\\ P_{(CA)} & P_{(CC)} & P_{(CG)} & P_{(CT)}\\ P_{(GA)} & P_{(GC)} & P_{(GG)} & P_{(GT)}\\ P_{(TA)} & P_{(TC)} & P_{(TG)} & P_{(TT)}\end{bmatrix} + \text{árvore}\right)$$

Modelo de substituição:

	A	C	G	T
A	$P_{AA}(t)$	$P_{AC}(t)$	$P_{AG}(t)$	$P_{AT}(t)$
C	$P_{CA}(t)$	$P_{CC}(t)$	$P_{CG}(t)$	$P_{CT}(t)$
G	$P_{GA}(t)$	$P_{GC}(t)$	$P_{GG}(t)$	$P_{GT}(t)$
T	$P_{TA}(t)$	$P_{TC}(t)$	$P_{TG}(t)$	$P_{TT}(t)$

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\nu/3}$$

probabilidade de mudança de estado

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\nu/3}$$

probabilidade de não mudança de estado

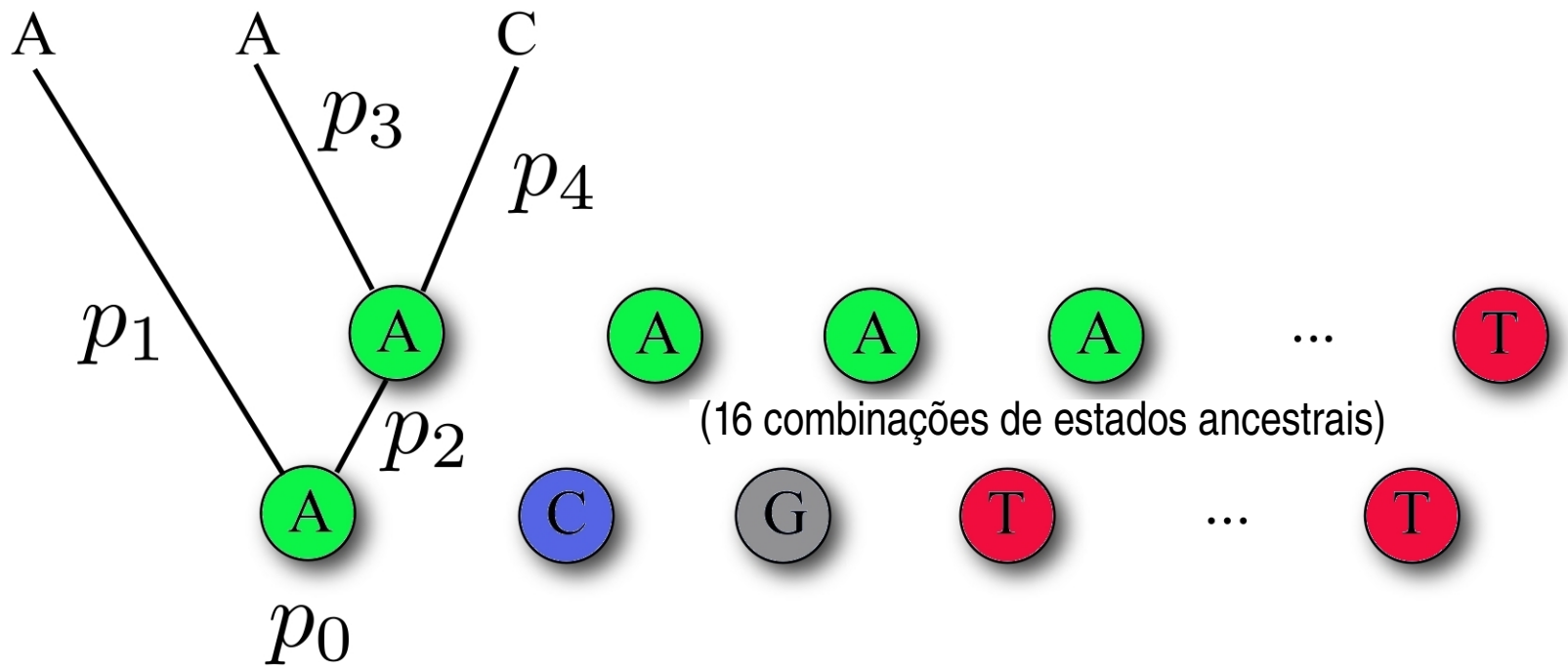
$v = 0.1$

	A	C	G	T
A	0.9064	0.0312	0.0312	0.0312
C	0.0312	0.9064	0.0312	0.0312
G	0.0312	0.0312	0.9064	0.0312
T	0.0312	0.0312	0.0312	0.9064

$v = 0.05$

	A	C	G	T
A	0.9516	0.0161	0.0161	0.0161
C	0.0161	0.9516	0.0161	0.0161
G	0.0161	0.0161	0.9516	0.0161
T	0.0161	0.0161	0.0161	0.9516

Lógica da inferência filogenética: Likelihood

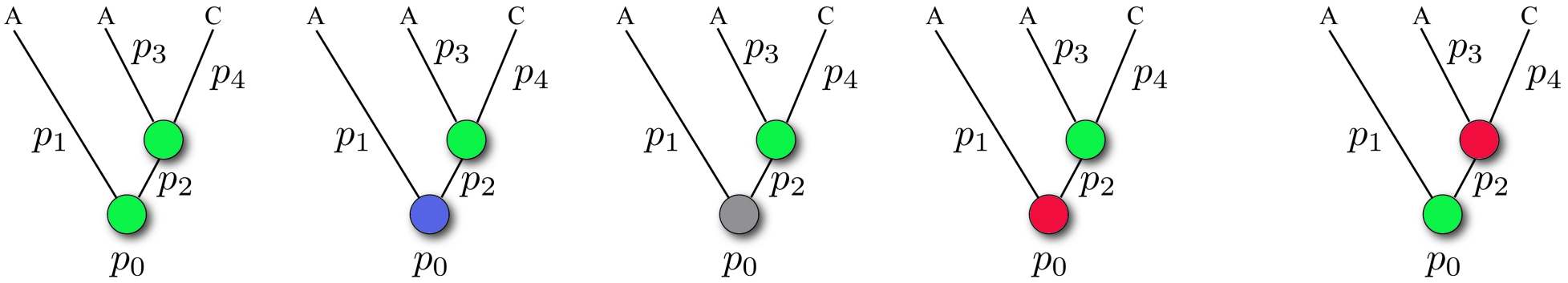


Regra 5: lintersecção de eventos, ou probabilidade conjunta: $P_{(A\ e\ B)} = P_{(A)} * P_{(B)}$.

$$Reconstrução\ 1 = P_0 * P_1 * P_2 * P_3 * P_4$$

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Lógica da inferência filogenética: Likelihood

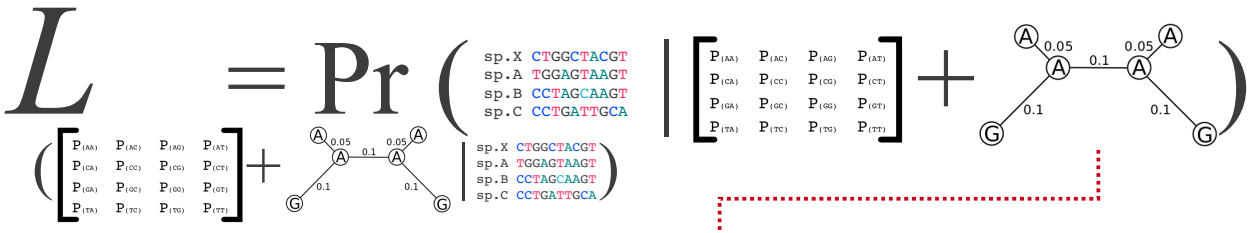


Reconstrução 1 + *Reconstrução* 2 + *Reconstrução* 3 + *Reconstrução* 4 + ... + *Reconstrução* 16

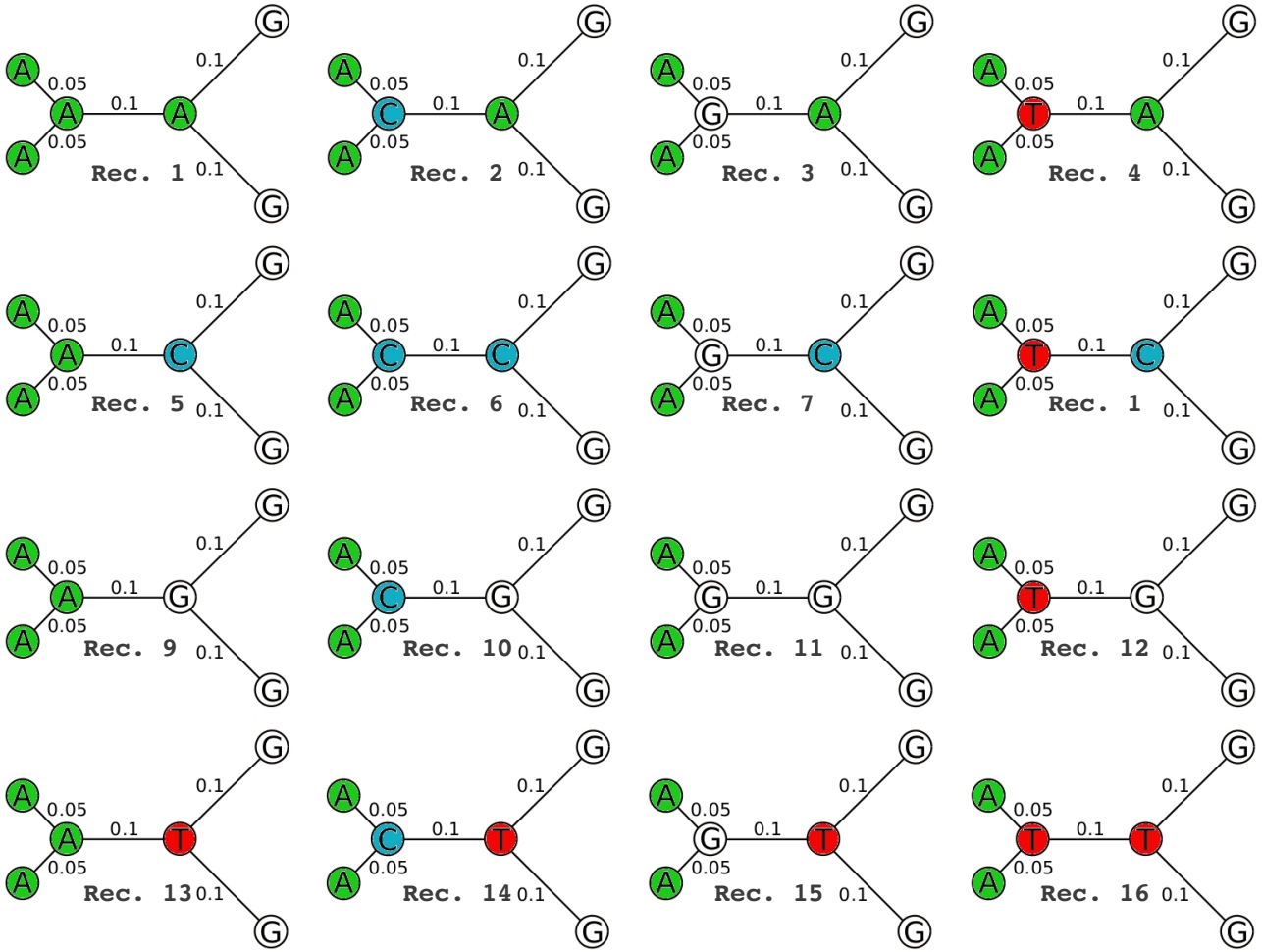
Regra 3: Eventos disjuntos: $P_{(A\ ou\ B)} = P_{(a)} + P_{(b)}$

Fonte: Modificado de Paul O. Lewis (UCONN), 2011.

Lógica da inferência filogenética: Likelihood



Avaliação de todas as reconstruções possíveis



Lógica da inferência filogenética: o cálculo

Para,

```
          1 2..  j          N
sp.X      CTGGCTA...CGT
sp.A      TGGAGTA...AGT
sp.B      CCTAGCA...AGT
sp.C      CCTGATT...GCA
```

A Verossimilhança Máxima de um **determinado caráter** (i.e., sítio) é:

$$L_{(j)} = P_{(rec.\ 1)} + P_{(rec.\ 2)} + P_{(rec.\ 3)} + \ldots + P_{(rec.\ n)}$$

A Verossimilhança Máxima (L^*) de uma **determinada hipótese** é dada por:

$$L = L_{(1)} * L_{(2)} * L_{(3)} * \ldots * L_{(n)} = \prod_{J=1}^{N} L_{(j)}$$

* tradicionalmente ela é avaliada pela soma dos logarítmos neperianos das probabilidades de cada caráter:

$$\ln L = \ln L_{(1)} + \ln L_{(2)} + \ln L_{(3)} + \ldots + \ln L_{(n)} = \sum_{J=1}^{N} \ln L_{(j)}$$

Lógica da inferência filogenética: modelos complexos

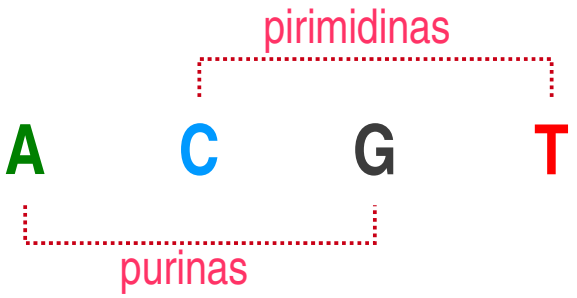
Considere:

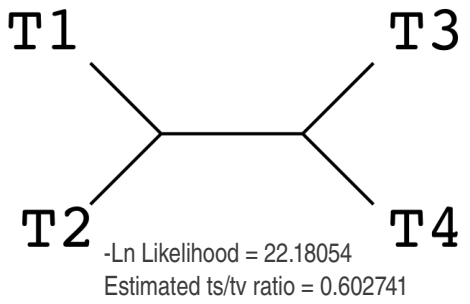
T1	A	C	G	T
T2	C	G	T	A
T3	G	T	A	C
T4	T	A	C	G

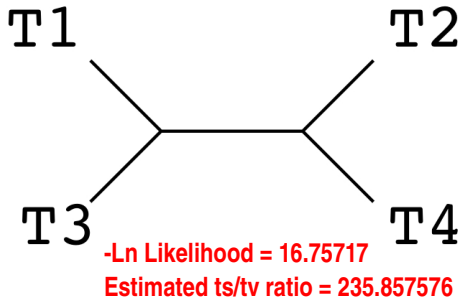
K2P

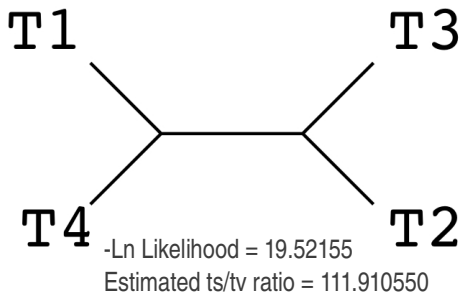
Kimura Two Parameters

	A	C	G	T
A	$P_{(A,A)}$	$P_{(A,C)}$	$P_{(A,G)}$ *k	$P_{(A,T)}$
C	$P_{(C,A)}$	$P_{(C,C)}$	$P_{(C,G)}$	$P_{(C,T)}$ *k
G	$P_{(G,A)}$ *k	$P_{(G,C)}$	$P_{(G,G)}$	$P_{(G,T)}$
T	$P_{(T,A)}$	$P_{(T,C)}$ *k	$P_{(T,G)}$	$P_{(T,T)}$









... onde *k* é a razão entre eventos de transição e transversão (valor estimado neste exemplo).

Note que utilizando o critério de parcimômia, nenhum caráter seria considerado informativo!

Seleção de Modelos

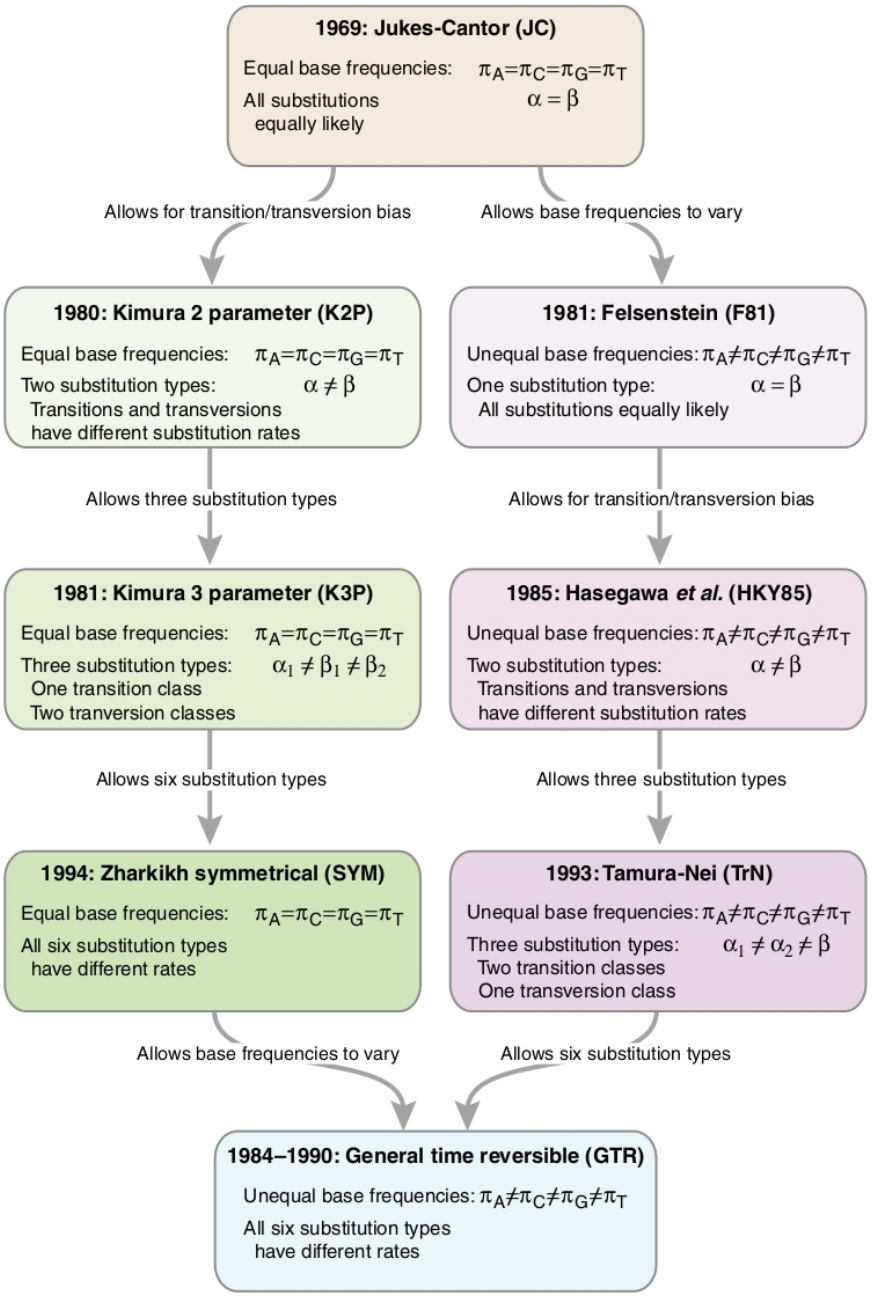
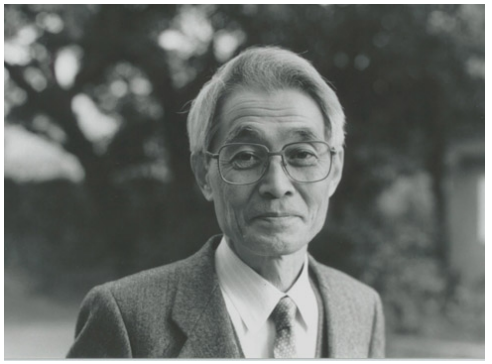


Figura: Simon et al. (2006)



"Model-selection theory began as a subject in statistics with Hirotugu Akaike's 1973 paper." (Sober, 2008:82)

Akaike's Information Criteria: "**AIC** provides a principled basis for deciding how fit-to-data should be traded off against simplicity." (Sober, 2008:86)

"As noted, the goal of AIC is to compare different models for their expected predictive accuracies." (Sober, 2008:92)

"The debate over AIC and BIC needs to be understood, in the first instance, as a debate over choice of goals – estimating predictive accuracy versus estimating average likelihood." (Sober, 2008:93)

Likelihood vs. Probabilidade:

Likelihood Criterion:

$$L_{(\text{mod.} \mid \text{obs.})} = \Pr(\text{observação} \mid \text{modelo})$$

Bayesian Criterion:

$$P_{(H|O)} = \frac{P_{(O|H)} \; P_{(H)}}{P_{(O)}}$$

- $P_{(O|H)}$: *Likelihood*
- $P_{(H)}$: *prior probability*
- $P_{(O)}$: *probabilidade incondicional da observação*