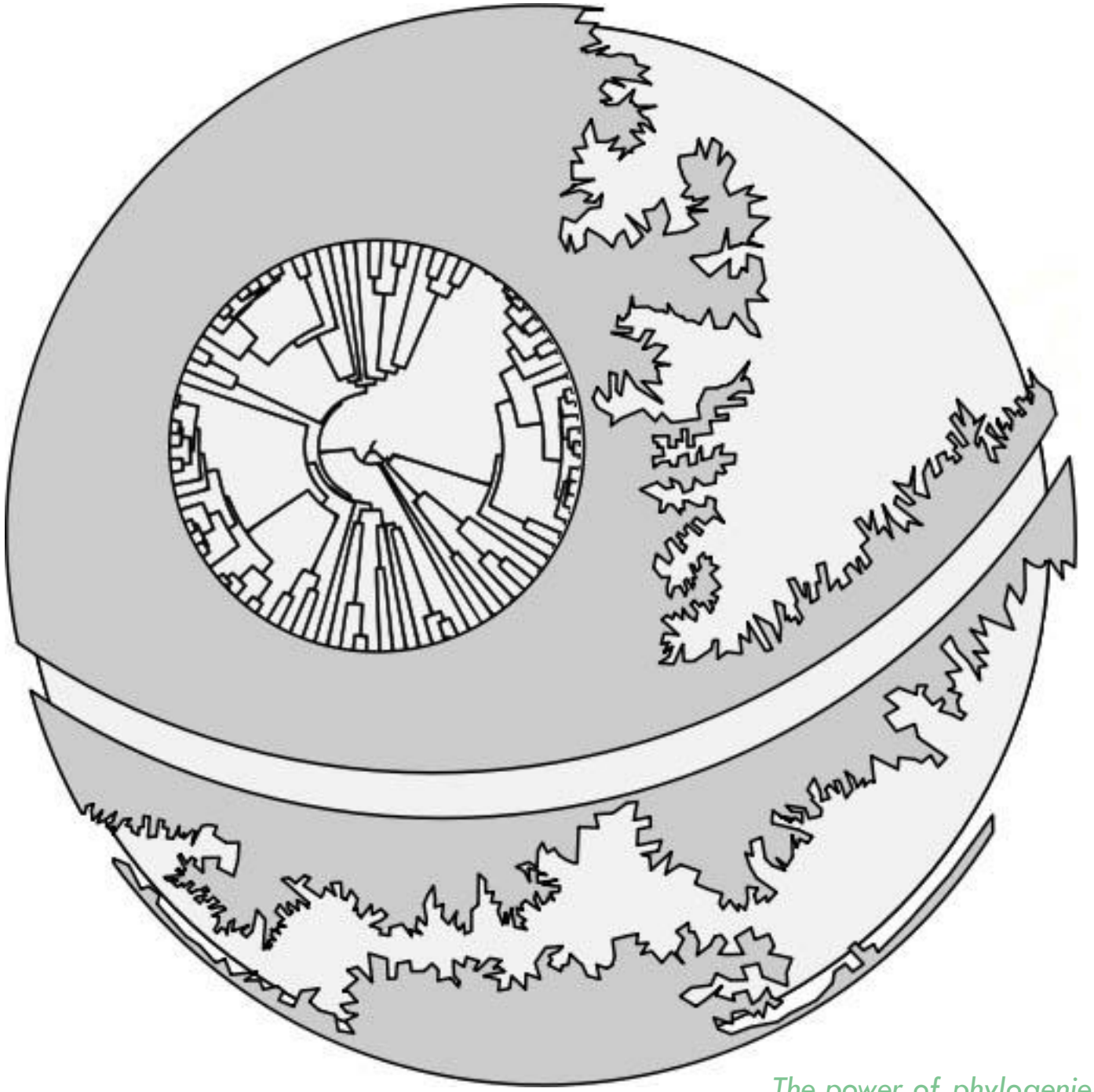
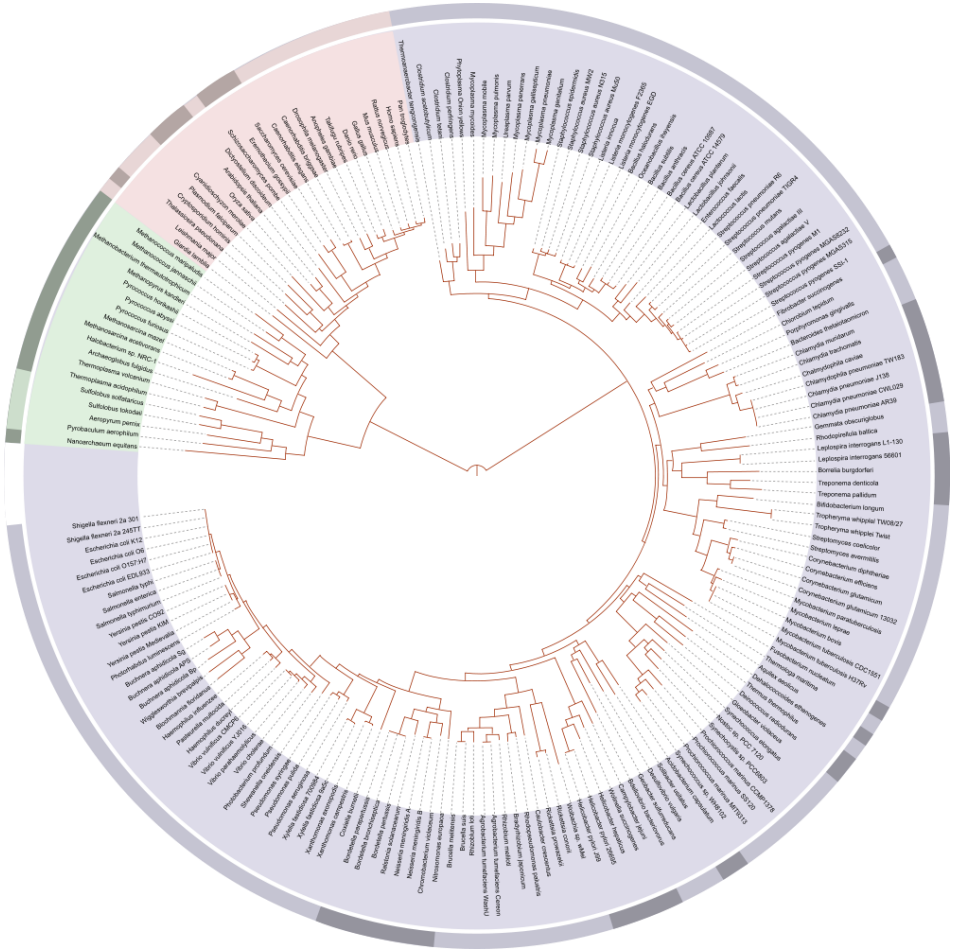
I CURSILLO DE FILOGENIA PARA USUARIOS

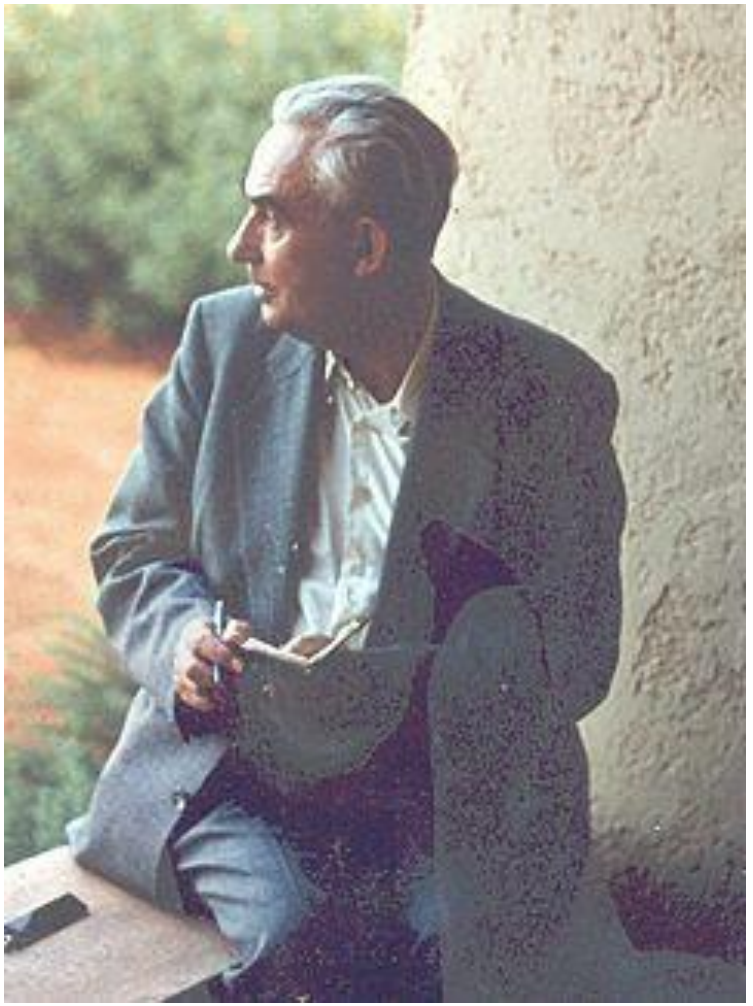
SESIÓN 1 – Recordatorio de conceptos básicos



The power of phylogenies, by Lanna Jin

WILLI HENNIG: FILOGENÉTICA





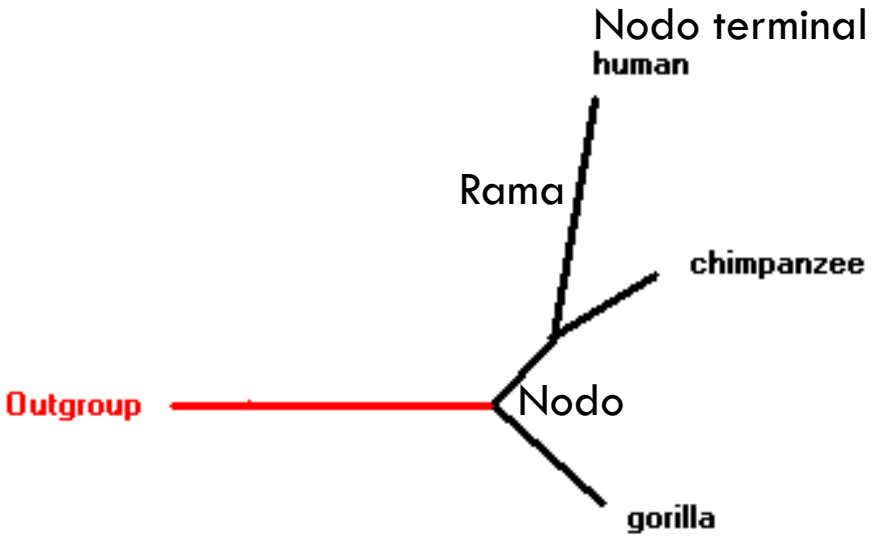
WILLI HENNIG: FILOGENÉTICA

"Grundzüge einer Theorie der phylogenetischen Systematik"

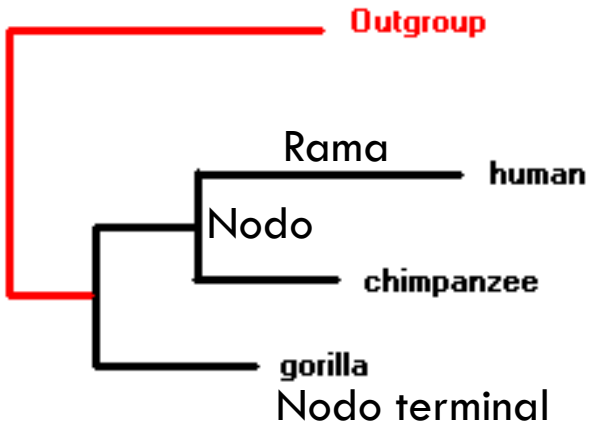
("Basic outline of a theory of phylogenetic systematics")

170 páginas, a lápiz

HABLANDO CON PROPIEDAD

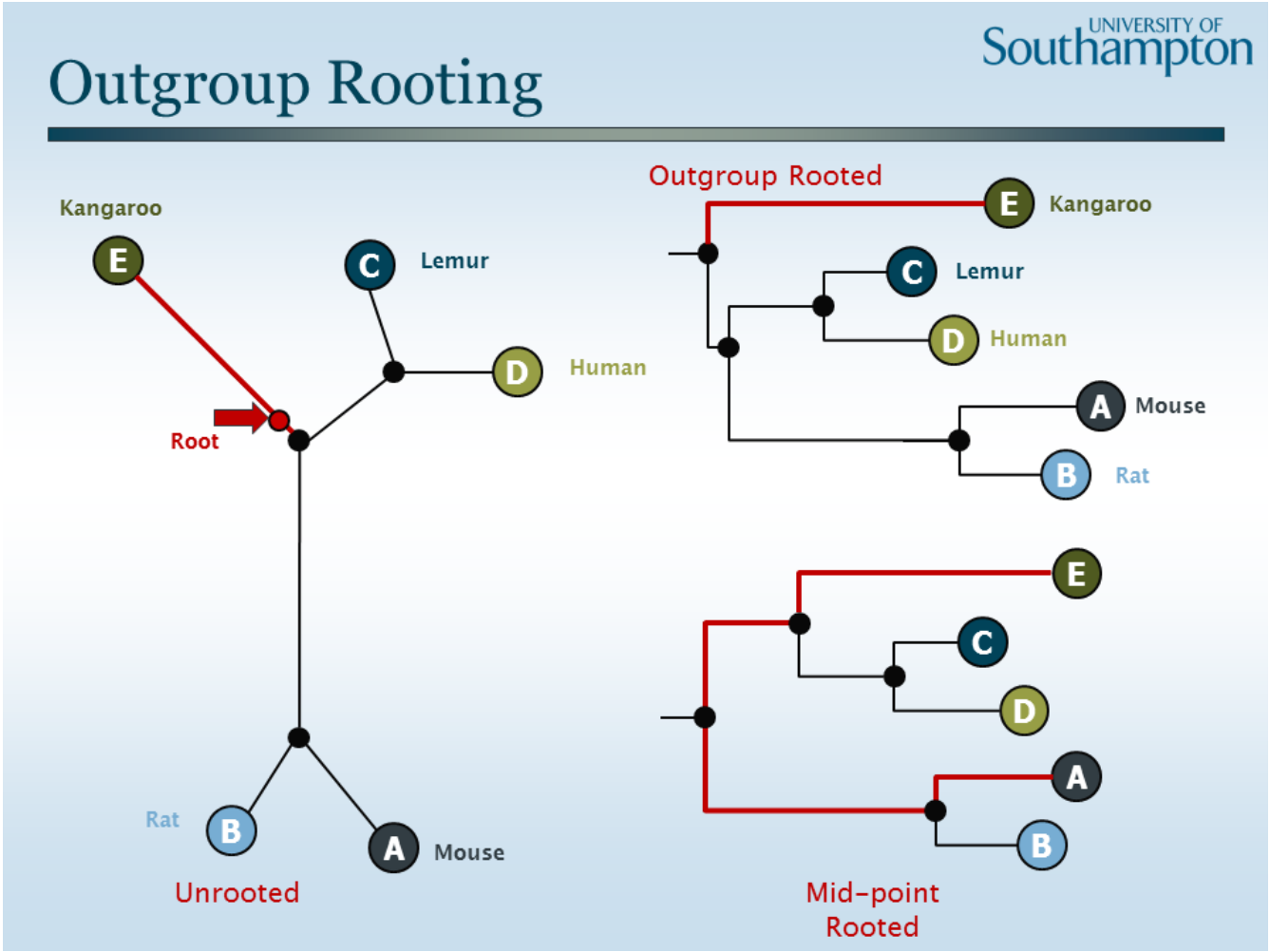


Árbol sin enraizar (*unrooted*)

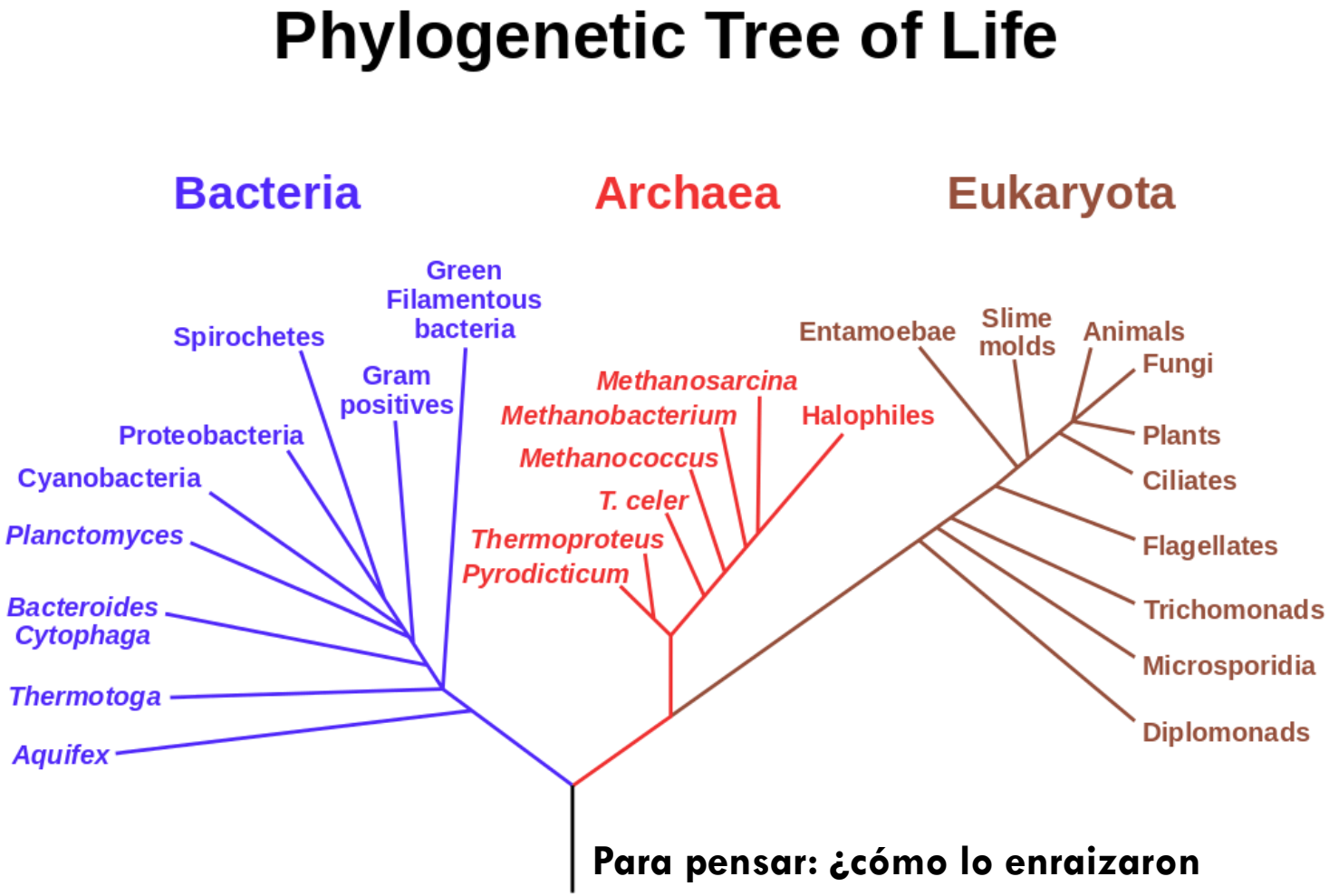


Árbol enraizado (*rooted*)

HABLANDO CON PROPIEDAD



HABLANDO CON PROPIEDAD



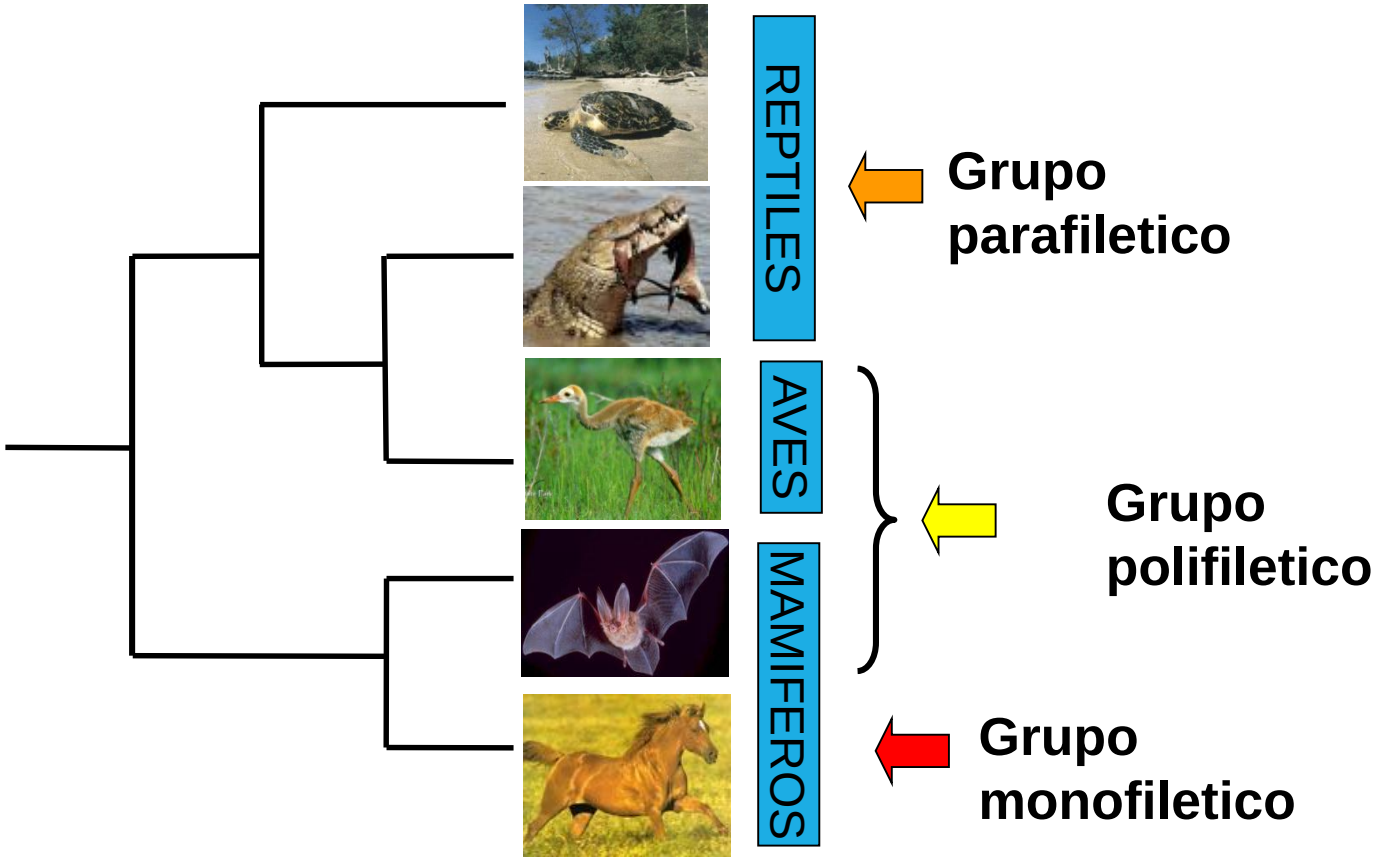
Para pensar: ¿cómo lo enraizaron si no existe outgroup?

HABLANDO CON PROPIEDAD

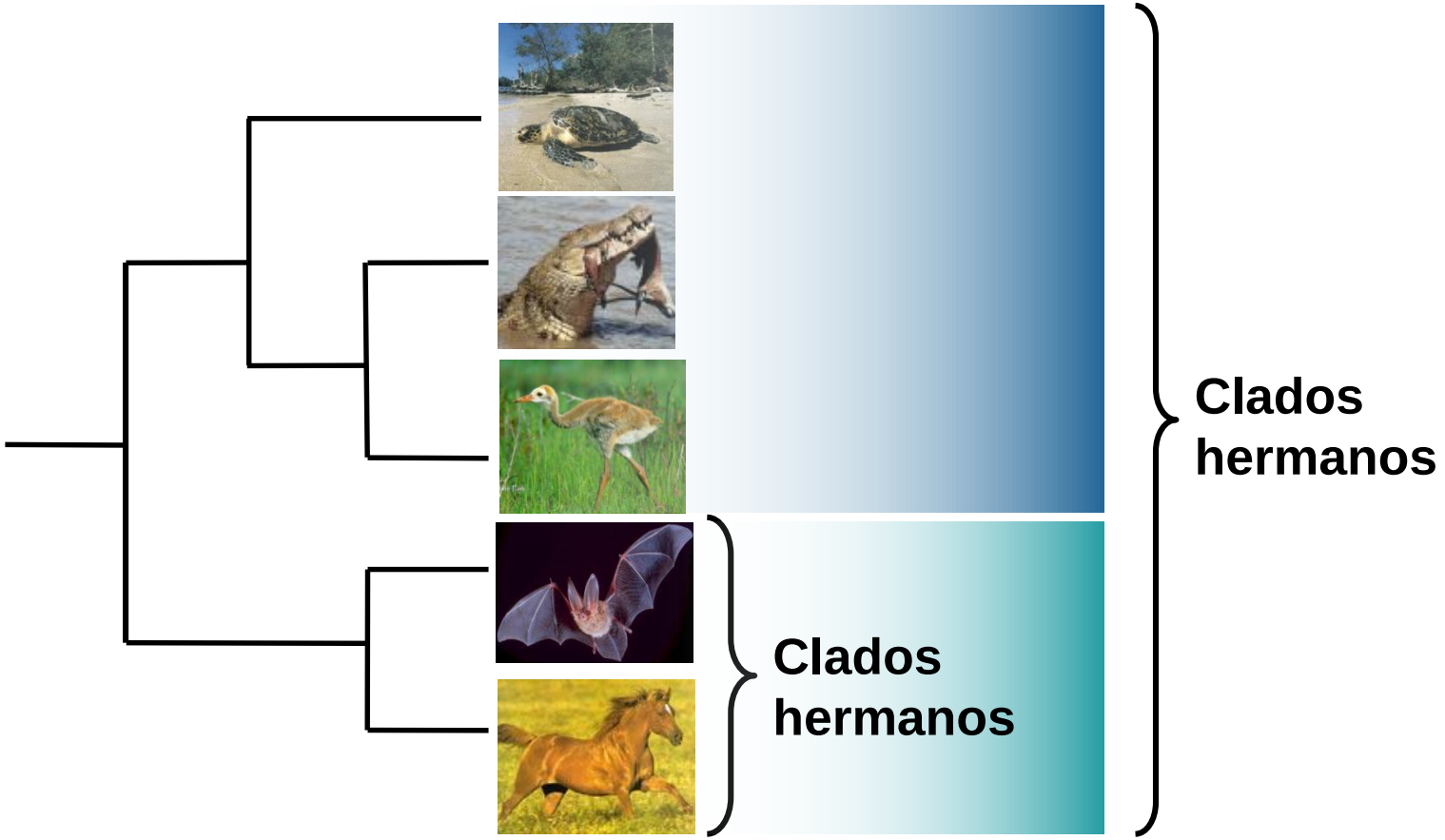
Tipos de árbol

- **Dendrograma:** Representación de un árbol filogenético (en general)
- **Cladograma:** Árbol obtenido por métodos cladísticos (MP). Sólo muestra información del patrón de ramificación.
- **Filograma:** Árbol en el que la longitud de las ramas es proporcional al número de sustituciones.
- **Cronograma:** Árbol en el que la longitud de las ramas es proporcional al tiempo transcurrido.

HABLANDO CON PROPIEDAD



HABLANDO CON PROPIEDAD



HABLANDO CON PROPIEDAD

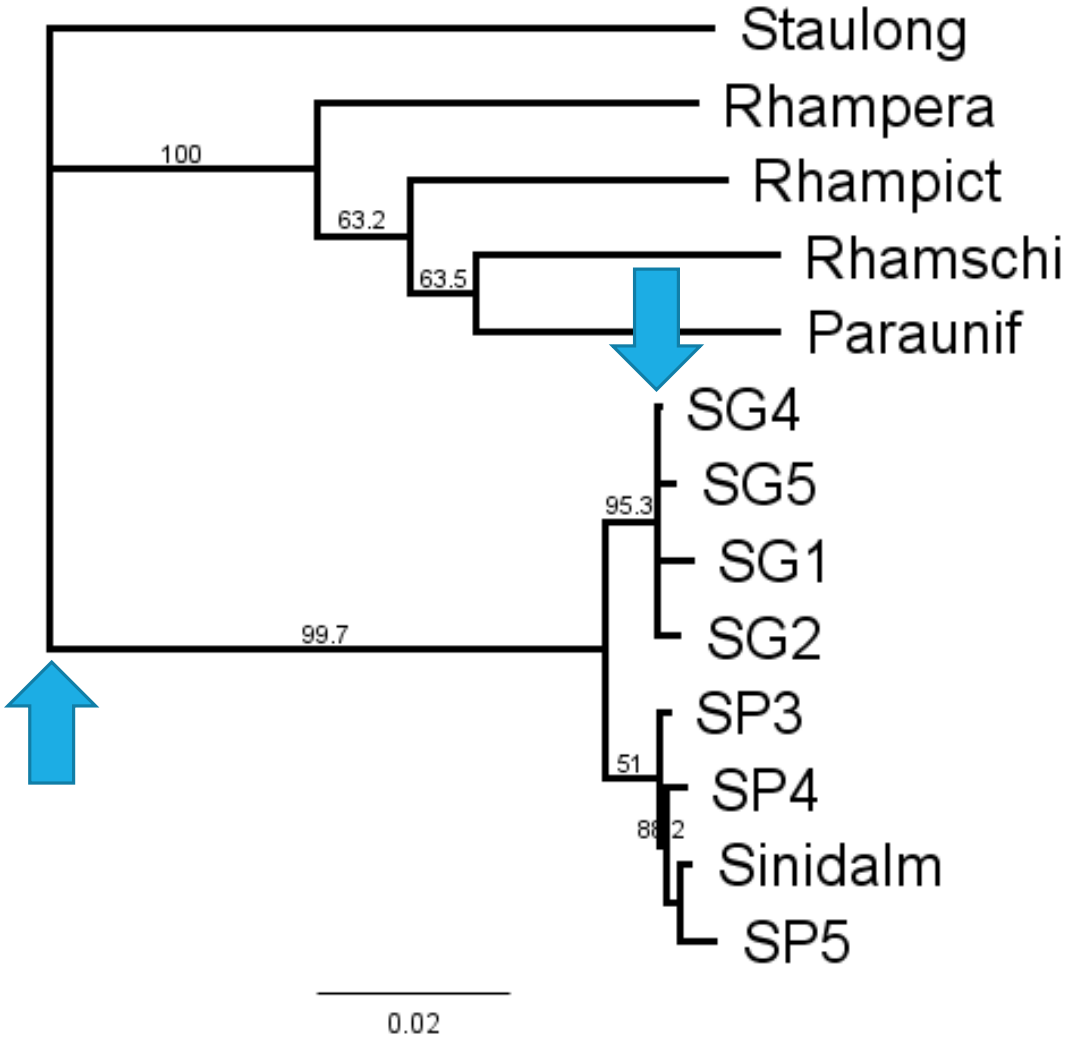
Politomías

Pocos datos

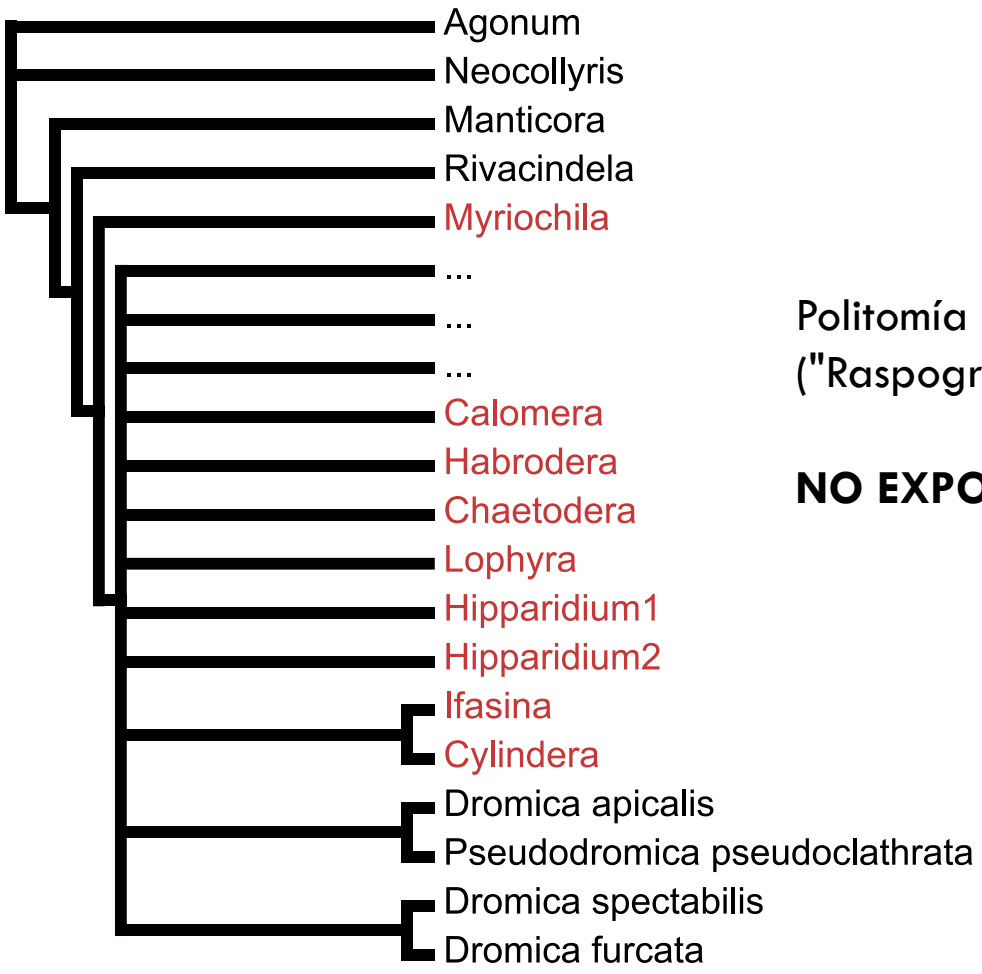
Malos datos

Secuencias idénticas

Consensos



HABLANDO CON PROPIEDAD



Politomía de categoría superior ("Raspograma")

NO EXPONER EN PÚBLICO

LOS DATOS

Formatos:

Phylip

FASTA

MEGA

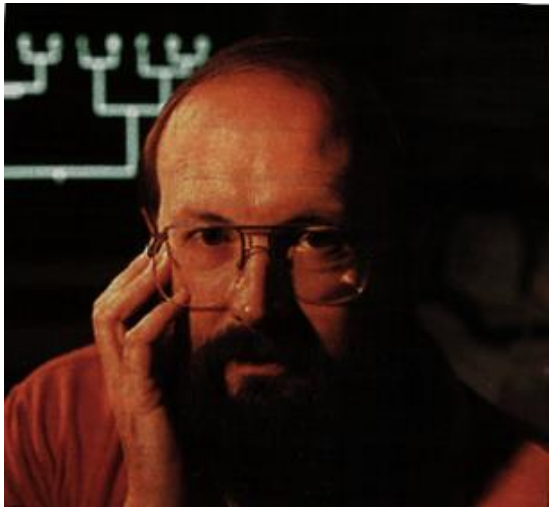
NEXUS

Newick

FORMATO PHYLIP

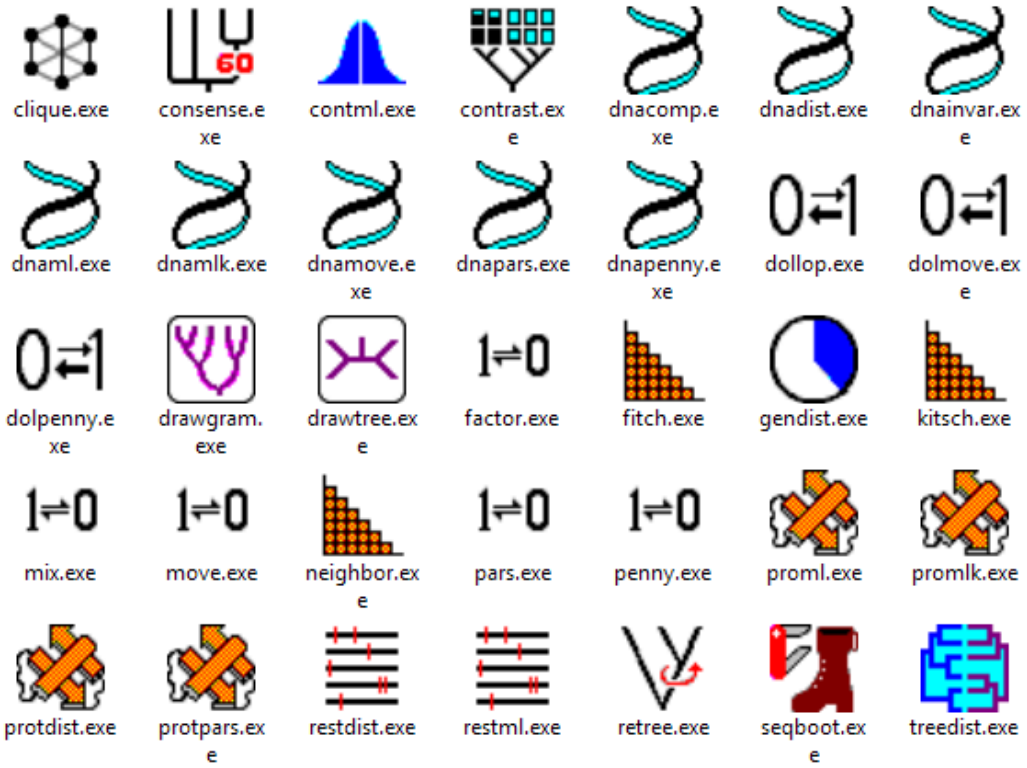
```
 8    6
Alpha1    AAGAAG
Alpha2    AAGAAG
Beta1     AAGGGG
Beta2     AAGGGG
Gamma1    AGGAAG
Gamma2    AGGAAG
Delta     GGAGGA
Epsilon   GGAAAG

**********NNNNNN
```

PREHISTORIA: PHYLIP



Joe Felsenstein 1980

PHYLogenetic Inference Package



FORMATO FASTA

Lipman & Pearson 1985

```
>Seq1
ATCGATCGGACGATCGATGCATCGACTG
>Seq2
AGCTAGCTACGATGCATTCGATCGATGCATCGATGC
>Seq3
ACGGACTCGTAGCAGCGACGGAGCATGCATCG
```

FORMATO FASTQ

Illumina 2009?

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Calidad (de menor a mayor):

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVW
XYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

FORMATO MEGA

Kumar, Tamura & Nei 1994

```
#Mega
Title: prueba

#168
ACTGATCGATCGATCGATGCACGCG
#169
GACTGATCGATGCTAGCTGACGATC
#171
ACGTAGCATGCTAGCTGATCATGCT
#176
GACTGACTGACTACTGCTGATGCTA
#199
ACGTCAGATGATCGATGTAGCATCG
```

FORMATO NEXUS

```
#NEXUS
begin assumptions;
charset 16s = 1-250;
charset coi = 251-808;
end;

begin data;
dimensions ntax=230 nchar=808;
format datatype=dna missing=? gap=-;
matrix
CibSLCU199  AAGGCTTAAATGTATGTAAAAGACGAGAAGACCCTATAGATCTTTATTTAATTAATATTTTGATAATTTA
CibSLCU200  AAGGCTTAAATGTATGTAAAAGACGAGAAGACCCTATAGATCTTTATTTAACTAATATTTTGATAATTTA
CibSCUE172  AAGGCTTAAATGTATGTAAAAGACGAGAAGACCCTATAGATCTTTATTTAACTAATATTTTGATAATTTA
ClaSCUN224  AAGGCTTAAATGTATGTAAAAGACGAGAAGACCCTATAGATCTTTATTTAACTAATATTTTGATAATTTA
ClaSCUN226  AAGGCTTAAATGTATGTAAAAGACGAGAAGACCCTATAGATCTTTATTTAACTAATATTTTGATAATTTA
ClaSCUN227  AAGGCTTAAATGTATGTAAAAGACGAGAAGACCCTATAGATCTTTATTTAACTAATATTTTGATAATTTA
CluPRMO41   AAGGCTTAAATGTATATAAAAGACGAGAAGACCCTATAGATCTTTATTTAATTAATATTTTGATAATTTA
CluPPTO79   AAGGCTTAAATGTATATAAAAGACGAGAAGACCCTATAGATCTTTATTTAATTAATATTTTGATAATTTA

...

Taiwan  AAGGCTTAAATGCTAATAATAGACGAGAAGACCCTATAGATCTTTATTTAATTAATATTTTAATAATTTAGGAT
duponti AAGGCTTAAATATTAATAATAGACGAGAAGACCCTATAGATCTTTATTTAATTAATATTTTAATAATTTAGGAT
;
end;
```

FORMATO NEXUS

Syst. Biol. 46(4):590–621, 1997

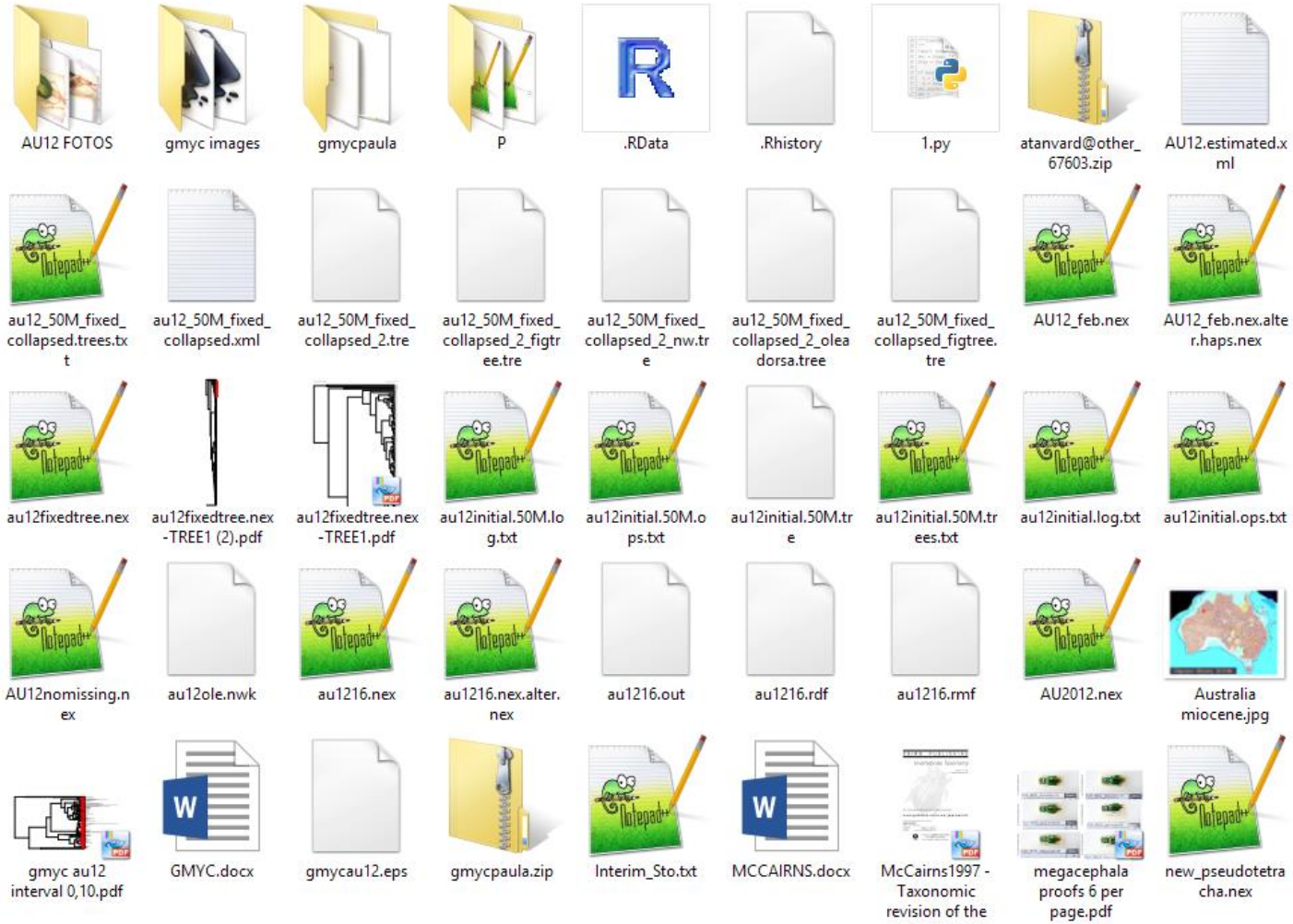
NEXUS: AN EXTENSIBLE FILE FORMAT FOR SYSTEMATIC INFORMATION

DAVID R. MADDISON,[1] DAVID L. SWOFFORD,[2] AND WAYNE P. MADDISON[3]

[1]*Department of Entomology, University of Arizona, Tucson, Arizona 85721, USA; E-mail: beetle@ag.arizona.edu*
[2]*Laboratory of Molecular Systematics, MRC 534, MSC, Smithsonian Institution, Washington, D.C. 20560, USA*
[3]*Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA*

Abstract.—NEXUS is a file format designed to contain systematic data for use by computer programs. The goals of the format are to allow future expansion, to include diverse kinds of information, to be independent of particular computer operating systems, and to be easily processed by a program. To this end, the format is modular, with a file consisting of separate blocks, each containing one particular kind of information, and consisting of standardized commands. Public blocks (those containing information utilized by several programs) house information about taxa, morphological and molecular characters, distances, genetic codes, assumptions, sets, trees, etc.; private blocks contain information of relevance to single programs. A detailed description of commands in public blocks is given. Guidelines are provided for reading and writing NEXUS files and for extending the format. [Computer program; file format; NEXUS; systematics.]

INCISO: CONSEJOS BIOINFORMÁTICOS



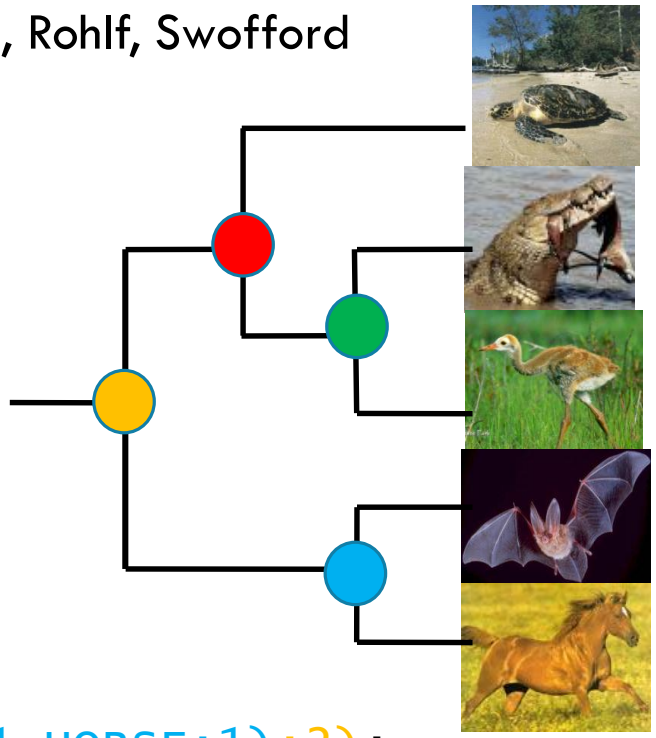
FORMATO NEWICK


```
((((((((((JP10_2:0.11896812086271513,JP10_3:0.11896812086271513):0.1783212341988487,
(((JP10_5:0.19832510284469906,JP10_4:0.19832510284469906):0.019102269901671853,JP10_
6:0.21742737274637092):0.04547329074948925,JP10_1:0.26290066349586017):0.03438869156
570368):0.3561342069722908,JP102_2:0.6534235620338547):0.8913034271965761,(((JP114_6
:0.18698705459923293,(JP114_2:0.023633231325136173,JP114_1:0.023633231325136173):0.1
6335382327409675):0.07342254236850332,((JP114_3:0.14300568834233562,JP114_4:0.143005
68834233562):0.05719461530753556,JP114_5:0.20020030364987118):0.060209293317865065):
0.8707521824641677,((JP112_3:0.19970440698103875,((JP112_5:0.062480880984406184,JP11
2_6:0.062480880984406184):0.05159420424420125,JP112_4:0.11407508522860743):0.0856293
2175243132):0.5789877485008885,(JP115_1:0.3061946444806791,(((JP115_5:0.070354479063
87304,JP115_6:0.07035447906387304):0.06821572346452509,JP115_4:0.13857020252839813):
0.1222191869128535,(JP115_3:0.08916610161244076,JP115_2:0.08916610161244076):0.17162
328782882087):0.045405255039427495):0.4724975110012481 5):0.3524696239499767):0.41356
520979852673):0.0897091795323064,((((JP112_1:0.037673099626655215,JP112_2:0.03767309
9626655215):0.36502250438963824,((JP102_3:0.03814615305933611,JP102_4:0.038146153059
33611):0.12669927686980564,JP102_5:0.16484542992914175):0.2378501740871517):0.027595
891721302834,JP102_1:0.4302914957375963):0.22535552590355357,(JP104_1:0.097485596434
28933,JP106_1:0.09748559643428933):0.5581614252068605):0.9787891471215873):0.9138000
996931765,(RV700:0.09476537062319945,RV699:0.09476537062319945):2.453470897832714):1
.2546290694895976,((((((JP12b_1:0.14934194034876924,((JP12b_4:0.07272317401878636,J
P12b_3:0.07272317401878636):0.0416901604089146,JP12b_2:0.11441333442770096):0.034928
60592106828):0.2704981588246449,JP12b_5:0.41984009917341414):1.0477102615452196,((((
JP12a_3:0.13299610615133872,(JP12a_2:0.10370451429679184,JP12a_5:0.10370451429679184
):0.029291591854546883):0.13633177872947133,JP12a_1:0.26932788488081005):0.402240199
9287393,(JP12a_6:0.22997905984052114,JP12a_4:0.22997905984052114):0.4415890249690282
):0.6889547588569944,(((JP13_4:0.16104863782360823,JP13_5:0.16104863782360823):0.510
9690211248816,JP13_6:0.6720176589484899):0.20272496241103344,((JP13_1:0.163951190419
08313,JP13_3:0.16395119041908313):0.18601332787621416,(JP13_2:0.2582940986388196,(JP
```

FORMATO NEWICK

Newick Restaurant (Dover, NH), 1986

Archie, Day, Felsenstein, Maddison, Meacham, Rohlf, Swofford



```
((TURTLE,(CROC,BIRD)),(BAT,HORSE));
```

```
((TURTLE:2,(CROC:1,BIRD:1):1),(BAT:1,HORSE:1):2);
```

LOS DATOS: LA GRAN CONTROVERSIA

Missing data

Datos que desconocemos para algunos taxones, debidos a secuencias ausentes o incompletas.

GAPs

"Saltos" en la secuencia debidos a inserciones o deleciones

```
format datatype=dna missing=? gap=-;
```

¡Ojo con algunos programas!

INCISO: MÉTODO "GEEK"

PERL

RUBY

python

INCISO: MÉTODO “GEEK”

```
'''
DEGAPPER
Escrito por Alejandro López, 18-06-2013
'''


def esgapterminal(i,line):
    if line[i] != '-':
        return False
    else:
        terminalpordelante = True
        terminalpordetras = True
        p = i - 1
        while terminalpordelante and (line[p] not in '\t\n '):
            if line[p] in 'ATCGWSMKRYBDHVNatcgwsmkrybdhvn':
                terminalpordelante = False
            p -= 1
        t = i + 1
        while terminalpordetras and (line[t] not in '\t\n '):
            if line[t] in 'ATCGWSMKRYBDHVNatcgwsmkrybdhvn':
                terminalpordetras = False
            t += 1
        if terminalpordelante or terminalpordetras:
            return True
        return False
```

```
infile = open('INFILE.nex','r')
outfile = open('OUTFILE.nex','w')

newfile = ''

for line in infile:
    if '\t' in line[1:]:
        newline = ''
        for i in range(len(line)):
            if esgapterminal(i,line):
                newline += '?'
            else:
                newline += line[i]
        line = newline
    newfile += line
outfile.write(newfile)
outfile.close()
infile.close()
```

LOS DATOS: LA GRAN CONTROVERSIA

El problema de los GAPs

```
ATCGA----AGCTAGGTAGCTAGCGATCGA
ATCGA----AGCTAGCTAGCTAGCGATCGA
ACCGA----AGCTAGCTAGCT-GCGATAGA
ATCGA--TGAGCTAGCTAGCT-GCGATCGA
ATCGA--TGATCTAGCTAGCT-GCTATCGA
ATCGATCTAGGCTAGCTAGCT-GAGATCGA
ATCGATCTGAGCTAGACAGCT-GCGATCGA
```

Soluciones:

A) Considerarlos como un quinto nucleótido (5th state): **NO**

B) Codificarlos como caracteres estándar: **PUEDE**

C) [WORK IN PROGRESS…]

I CURSILLO DE FILOGENIA PARA USUARIOS

SESIÓN 2 – Distancias y parsimonia

UPGMA: DISTANCIAS A LO BRUTO

Sokal & Michener 1958

Reconstrucción jerárquica

Se calculan las distancias entre cada secuencia y las demás; por ejemplo, mediante el índice de Jaccard:

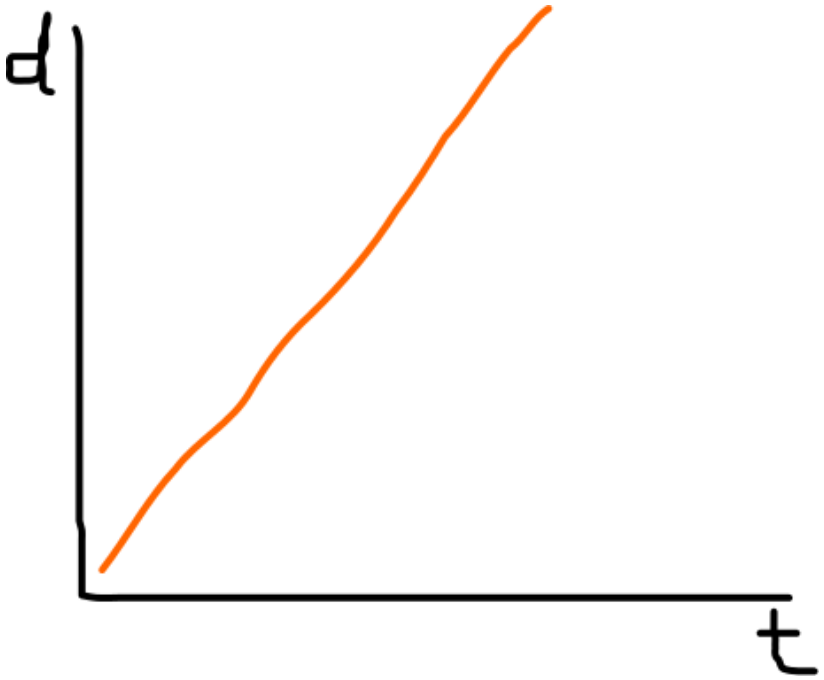
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Se va dibujando el árbol uniendo las más próximas (menor distancia)

http://www.southampton.ac.uk/~re1u06/teaching/upgma/

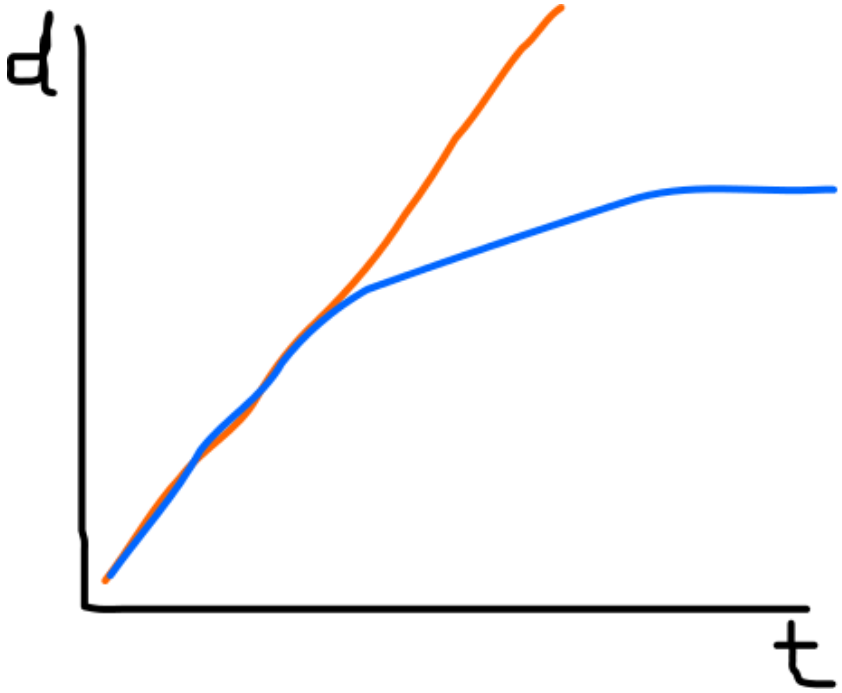
UPGMA: DISTANCIAS A LO BRUTO

Sería perfecto si distancia = diferencias = tiempo

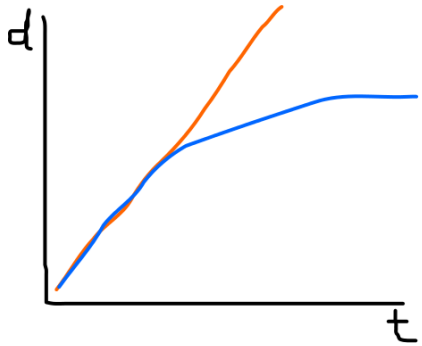


UPGMA: DISTANCIAS A LO BRUTO

Pero…

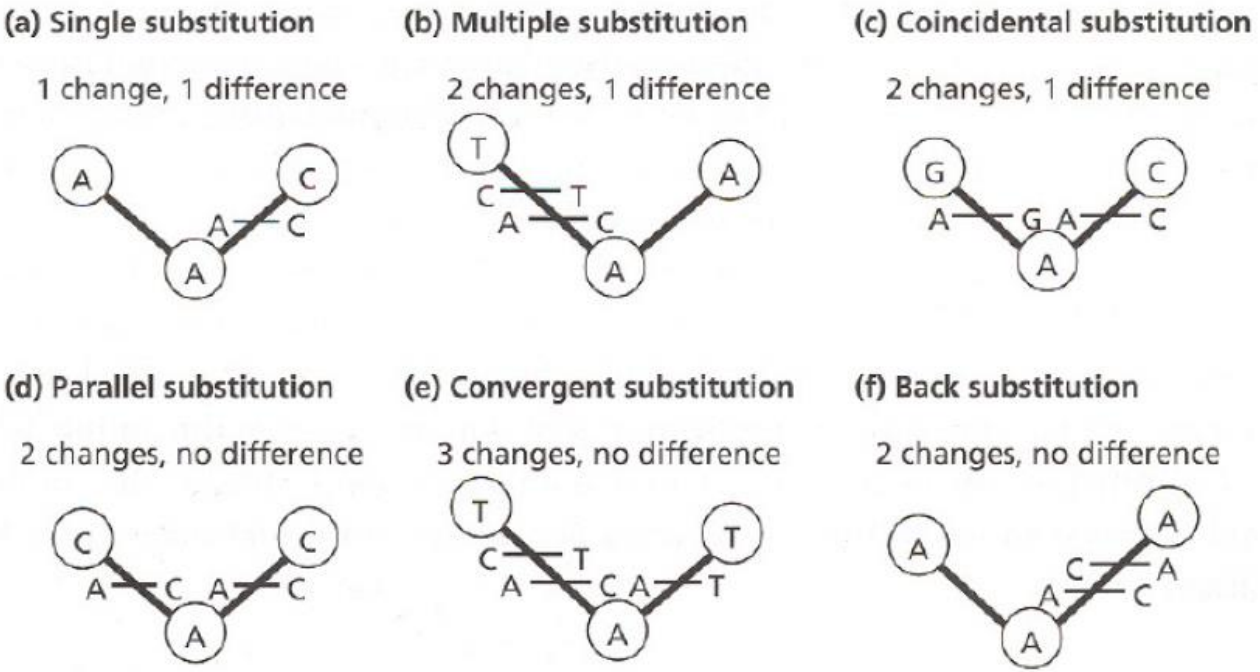


UPGMA: DISTANCIAS A LO BRUTO

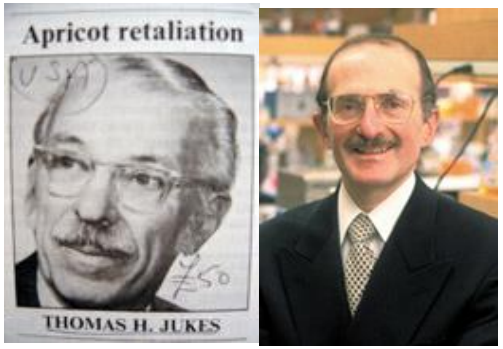


¿Por qué?

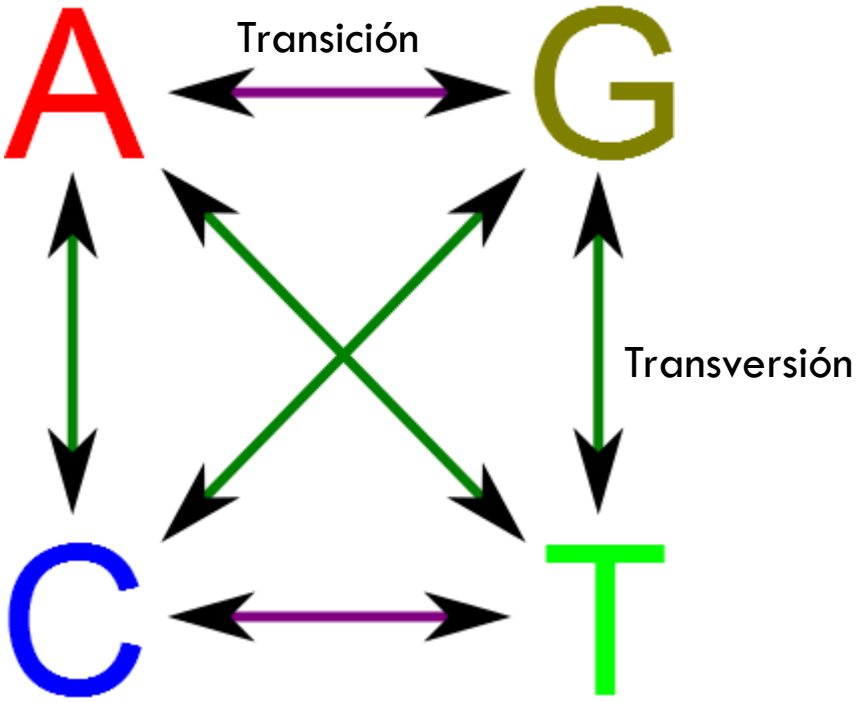
Porque 1 diferencia ≠ 1 sustitución



MODELOS DE SUSTITUCIÓN



Jukes & Cantor 1969



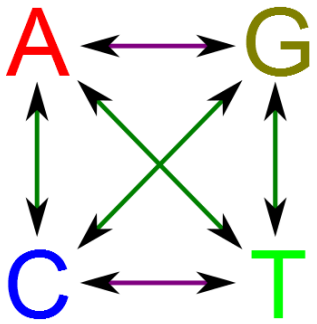
$$P_i = \begin{bmatrix} . & \alpha & \alpha & \alpha \\ \alpha & . & \alpha & \alpha \\ \alpha & \alpha & . & \alpha \\ \alpha & \alpha & \alpha & . \end{bmatrix}$$

$$f = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

MODELOS DE SUSTITUCIÓN

Kimura 1980: Transiciones ≠ Transversiones

$$P_i = \begin{bmatrix} . & \beta & \alpha & \beta \\ \beta & . & \beta & \alpha \\ \alpha & \beta & . & \beta \\ \beta & \alpha & \beta & . \end{bmatrix}$$



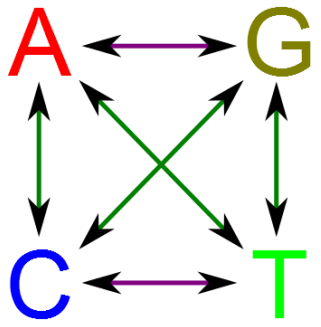
$$f = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

MODELOS DE SUSTITUCIÓN

Felsenstein 1981: Diferentes frecuencias

$$P_i = \begin{bmatrix} . & \pi_C\alpha & \pi_G\alpha & \pi_T\alpha \\ \pi_A\alpha & . & \pi_G\alpha & \pi_T\alpha \\ \pi_A\alpha & \pi_C\alpha & . & \pi_T\alpha \\ \pi_A\alpha & \pi_C\alpha & \pi_G\alpha & . \end{bmatrix}$$

$$f = \begin{bmatrix} \pi_A & \pi_C & \pi_G & \pi_T \end{bmatrix}$$



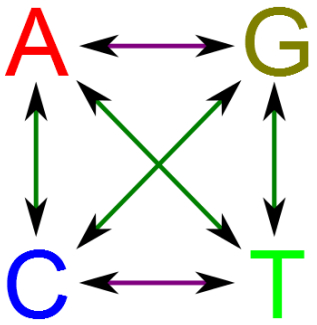
MODELOS DE SUSTITUCIÓN

Tasas de sustitución	Frecuencias iguales	Frecuencias diferentes
1	Jukes-Cantor (JC)	Felsenstein (F81)
2	Kimura 2-parámetros (K2P, K80)	HKY85
3	K3ST y TrNef	K81uf y TrN
4	TIMef	TIM
5	TVMef	TVM
6	SYM	GTR

MODELOS DE SUSTITUCIÓN

GTR (Tavaré, 1986)

$$P_i = \begin{bmatrix} . & \pi_C\alpha & \pi_G\beta & \pi_T\gamma \\ \pi_A\alpha & . & \pi_G\delta & \pi_T\varepsilon \\ \pi_A\beta & \pi_C\delta & . & \pi_T\theta \\ \pi_A\gamma & \pi_C\varepsilon & \pi_G\theta & . \end{bmatrix}$$



$$f = [\pi_A \quad \pi_C \quad \pi_G \quad \pi_T]$$

NEIGHBOR-JOINING

Saitou & Nei 1987

Las distancias se corrigen y el árbol se construye a partir de una **matriz Q** que se calcula así:

$$Q(i,j) = (r-2)d(i,j) - \sum_{k=1}^{r} d(i,k) - \sum_{k=1}^{r} d(j,k)$$

NEIGHBOR-JOINING

Ventajas

Rápido

No requiere gran potencia de cálculo

Inconvenientes

Muy poco fiable

Suelen aparecer artefactos

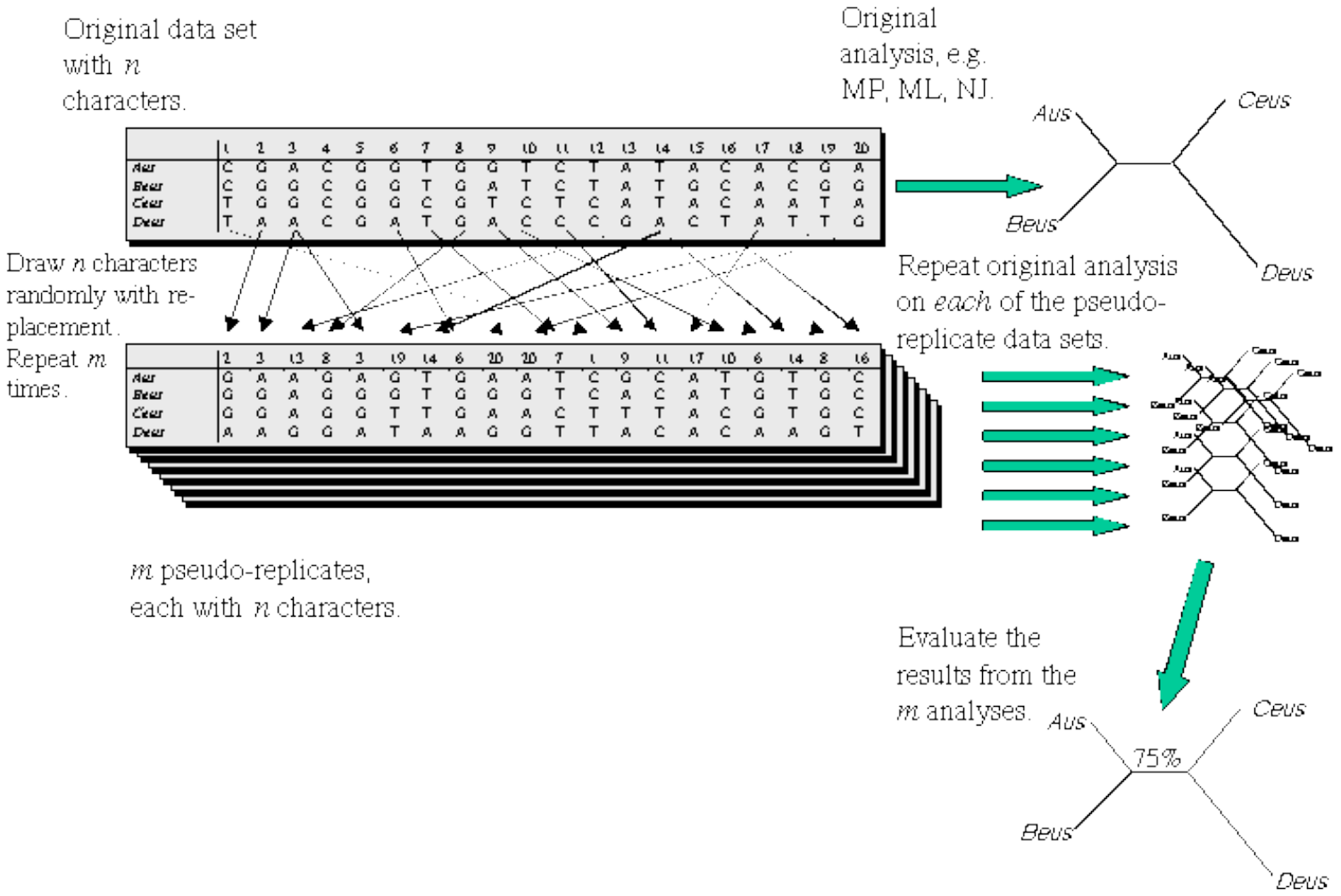
BOOTSTRAP

Para averiguar cómo de fiables son los nodos que obtenemos.

Es un método de “resampling”: se genera una serie de réplicas de la matriz original, se repite el análisis en cada una de ellas, y al final, para cada nodo del árbol se ve en qué porcentaje de estas réplicas se ha recuperado dicho nodo.

Las matrices de réplica se construyen tomando nucleótidos al azar de la matriz original.

BOOTSTRAP


Original data set with *n* characters.
Original analysis, e.g. MP, ML, NJ.
Aus
Ceus
Beus
Deus
Draw *n* characters randomly with replacement. Repeat *m* times.
Repeat original analysis on *each* of the pseudo-replicate data sets.
m pseudo-replicates, each with *n* characters.
Evaluate the results from the *m* analyses.
75%

NEIGHBOR-JOINING

geneious

MÁXIMA PARSIMONIA

Navaja de Occam / Lex parsimoniae (Guillermo de Ockham, s.XIV):

"Las entidades no deben multiplicarse más allá de lo necesario"

"En igualdad de condiciones, la explicación más sencilla suele ser la correcta"

MÁXIMA PARSIMONIA

Ley de Malcolm:

"La vida se abre camino"

MÁXIMA PARSIMONIA

1. Mide la longitud de todos los árboles posibles
2. El árbol con menor longitud será el correcto

Cómo medir la longitud:

-Para cada carácter de la matriz:

-Para cada nodo del árbol:

-Si se detecta que ha habido un cambio, sumarlo al total.

MÁXIMA PARSIMONIA

El problema: el número de árboles posibles

$$N = (2T - 3) \prod_{i=3}^{T} (2i - 5)$$

2 taxones → 1 árbol

3 taxones → 3 árboles

5 taxones → 105 árboles

10 taxones → 34.459.425 árboles

50 taxones → $6 \cdot 10^{81}$ árboles > Número de átomos en el Universo

MÁXIMA PARSIMONIA

Tipos de búsqueda:

A) Exhaustiva

B) Branch & Bound

C) Heurística

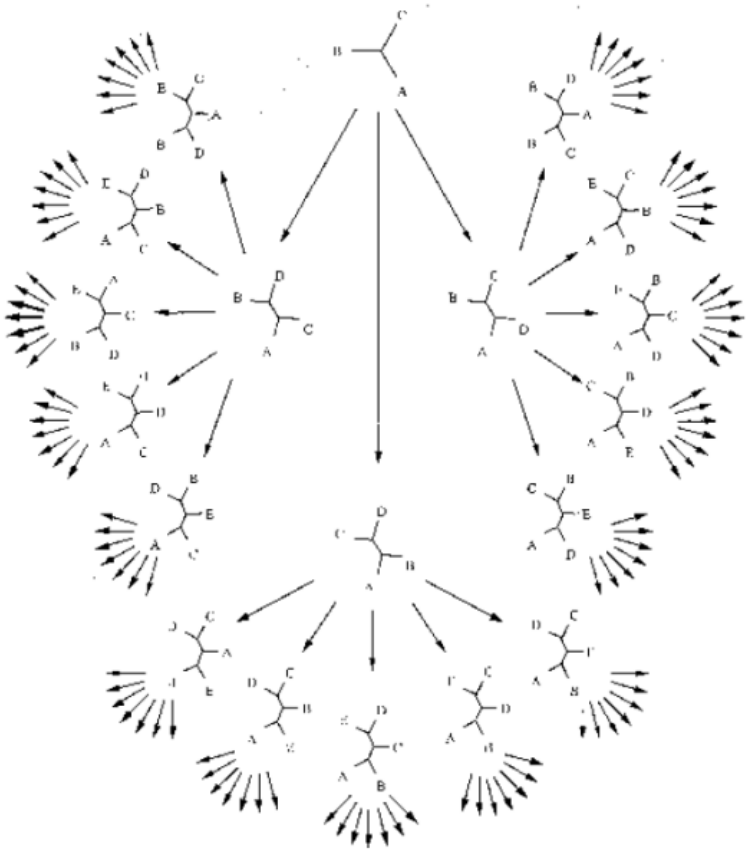
MÁXIMA PARSIMONIA

Exhaustiva

Va añadiendo taxones uno a uno y midiendo la longitud de cada árbol

Siempre encuentra el árbol más corto

Mide TODOS los posibles árboles

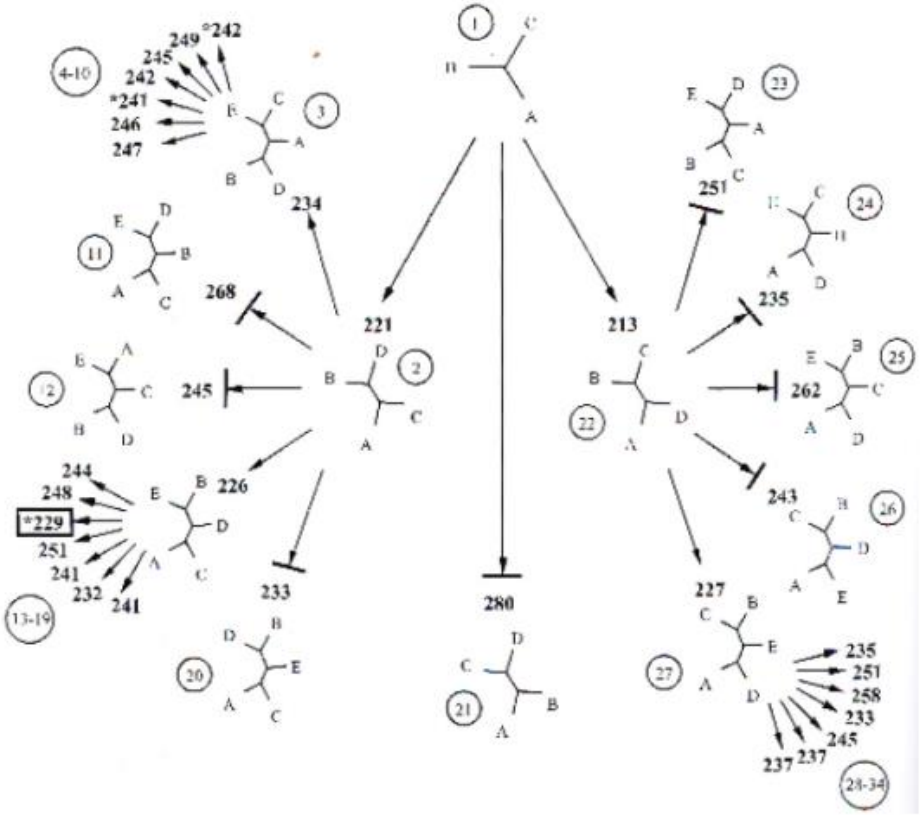


MÁXIMA PARSIMONIA

Branch & Bound

Como la exhaustiva: va añadiendo taxones uno a uno en cada paso mide todos los árboles que va obteniendo, pero sólo pasan al siguiente paso los que no superan un umbral de longitud.

Garantiza encontrar el más corto.



MÁXIMA PARSIMONIA

Heurística

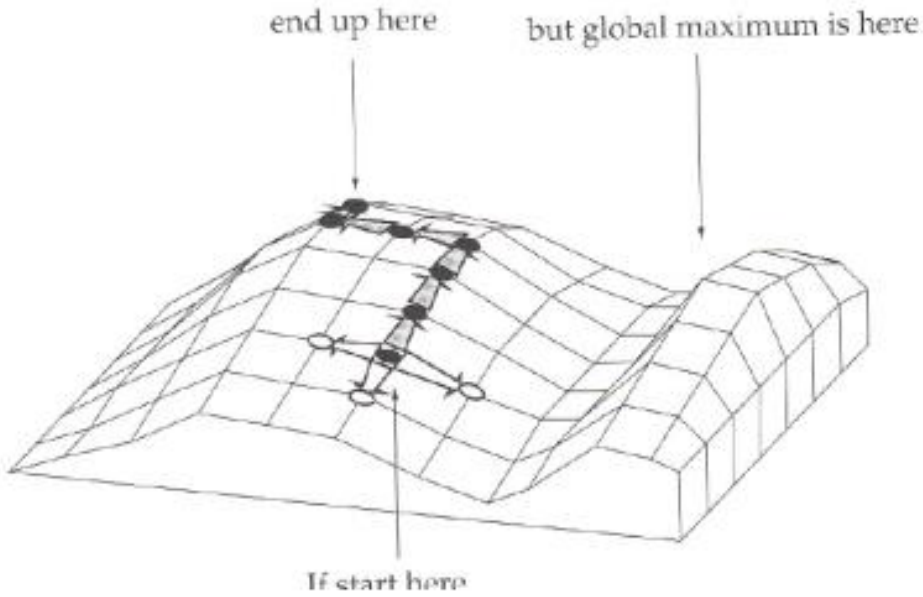
Varios métodos:

- Stepwise addition
- Star decomposition

Branch swapping:

- NNI: Nearest neighbor interchange
- SPR: Subtree pruning and regrafting
- TBR: Tree bisection and reconnection

Es rápida, pero puede equivocarse



MÁXIMA PARSIMONIA

¡Peligro! Fenómeno de atracción de ramas largas

Las ramas que evolucionan rápidamente tienden a situarse juntas, aunque no tengan ninguna relación real entre ellas.

Ejemplos clásicos:

- El clado de “mamíferos africanos”
- “La cobaya no es un roedor”
- “El erizo es un mamífero ancestral”
- Avispas de las agallas
- Microsporidios