Considérations sur l'analyse des données dans le cadre du barcode ADN

Frederic Austerlitz[1], Olivier David[2], Brigitte Schaeffer[2], Kevin Bleakley[5,6,7], Madalina Olteanu[2], Raphael Leblois[3], Michel Veuille[3,4], Catherine Laredo[2,8].

[1]CNRS, Laboratoire Ecologie Systématique et Evolution, UMR 8079, Orsay, F-91405; Univ Paris-Sud, Orsay, F-91405; AgroParisTech, Paris, F-75231, France.
[2]UR341, Mathématiques et informatique appliquées, INRA, F-78350 Jouy-en-Josas, France.
[3]Muséum National d'Histoire Naturelle, UMR 5202 MNHN/CNRS, Laboratoire Origine, Structure et Evolution de la Biodiversité, 16 rue Buffon, 75005 Paris, France.
[4]Laboratoire de Biologie intégrative des populations, Ecole Pratique des Hautes Etudes, Paris, France.
[5]Institut Curie, Centre de Recherche, Paris, F-75248 France.
[6]INSERM, U900, Paris, F-75248 France.
[7]Centre for Computational Biology, Ecole des Mines de Paris, 35 rue St Honoré, Fontainebleau, F-77305 France.
[8]Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 et 7, UMR CNRS 7599, 4 place Jussieu, 75005 Paris, France.

Aims of DNA barcoding

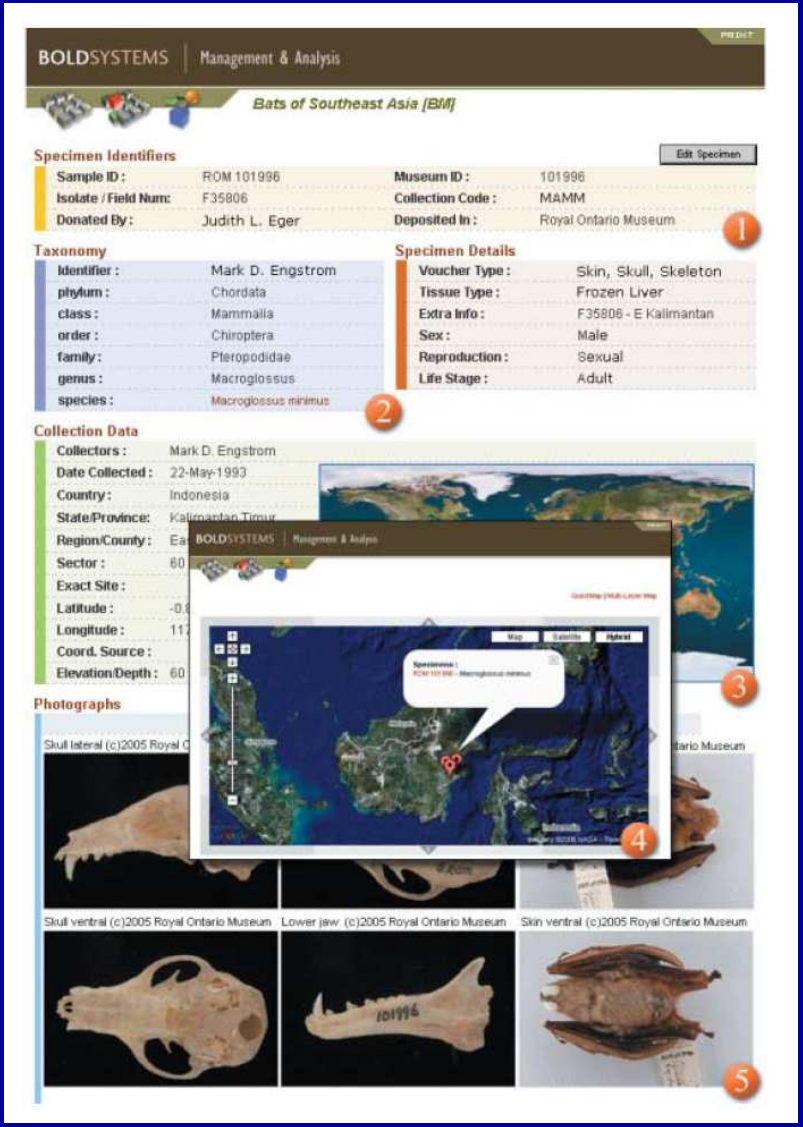
- Assign individuals to a given species according to sequence at a given locus
  - Degraded samples.
  - Larvae
  - Very similar species
- Identify criptic species from these DNA sequences.

Barcoding of life initiative

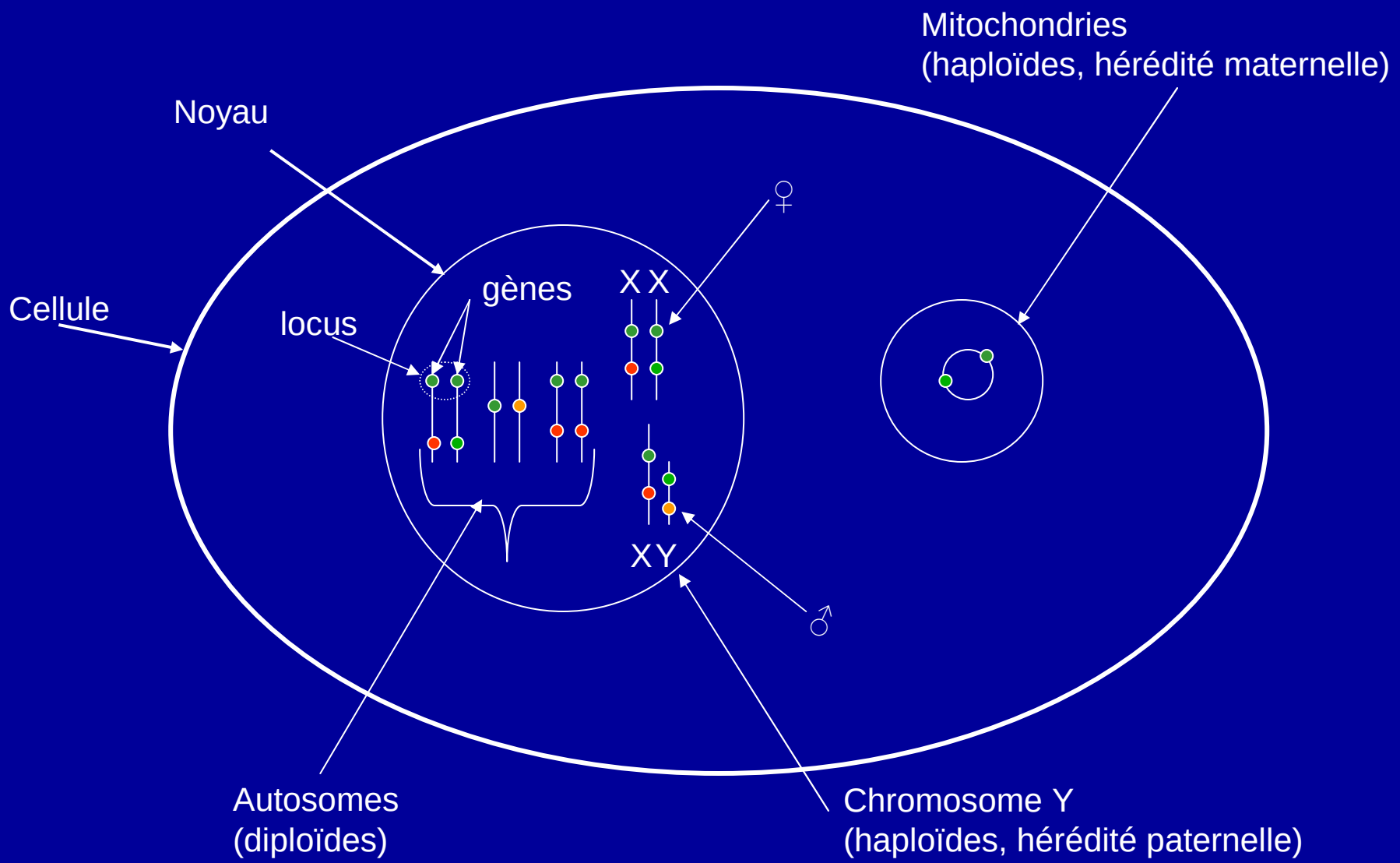
- Develop a reference library containing sequences for one mitochondrial gene: cytochrome *c* oxidase I (COI) with individuals with known status.
- Several sample per species (from 5 to 30)
- Assign individuals of unknown taxonomic status to given species by comparing their sequence with the sequence present in the database.

  ↳What is the best assignation method?

Example of data available



Différents génomes chez un même individu. Animaux (cas général)



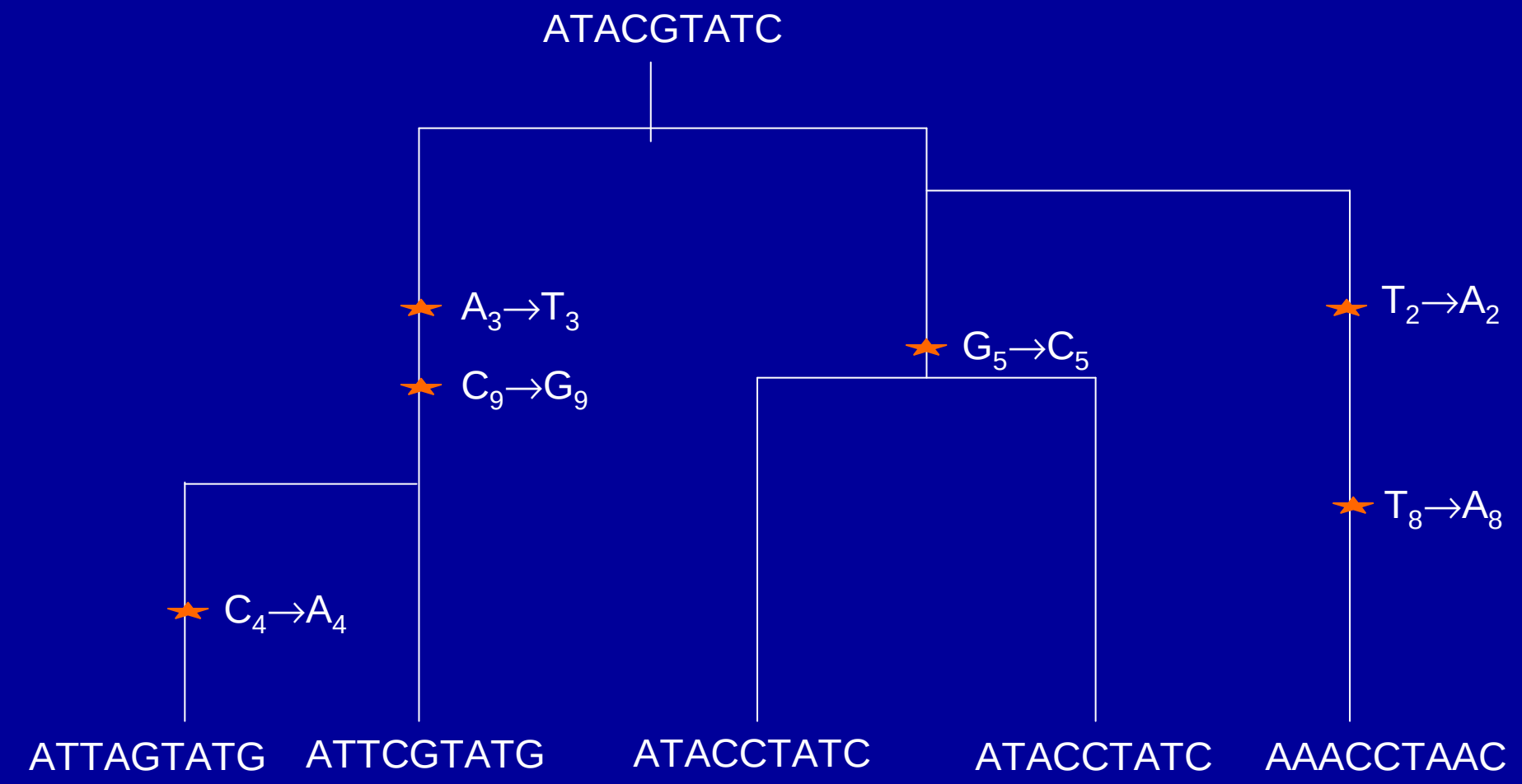
SEQUENCES

- Echantillon de séquences prises chez plusieurs individus pour un même locus donné

Ecotype‡	311*§	426	461*	704*	1092	1233	1245	1255*	1311	1315*	1326	1359	1374	1438	1440	1458	1543*	1548
Col-0¶ (R)	A	A	C	G	A	C	A	T	C	C	C	T	T	T	G	G	G	C
Ler-0 (R)	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Tsu-0 (R)	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.
Ws-0 (P)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
Wa-0 (S)	G	C	.	.	T	.	.	C	T	A	.	C	C	C	A	A	.	T
Zu-0 (S)	G	C	.	.	T	T	T	C	T	A	.	C	C	C	A	A	.	T
Zu-0-1 (S)	G	C	.	.	T	T	T	C	T	A	.	C	C	C	A	A	.	T
Zu-0-3 (S)	G	C	.	.	T	T	T	C	T	A	.	C	C	C	A	A	.	T
Zu-0-4 (S)	G	C	.	.	T	T	T	C	T	A	.	C	C	C	A	A	.	T
Zu-0-6 (S)	G	C	.	.	T	T	T	C	T	A	.	C	C	C	A	A	.	T
Zu-0-7 (S)	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Zu-0-8 (S)	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
UIE132 (P)	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.
rps2-210C(S)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

- Longueur typique : jusqu'à quelques kb.
- Taux de mutation usuel : 10^{-8} à 10^{-7} par nucléotide par génération.

La divergence entre les séquences reflète leur temps de coalescence



Testing the assignation methods

- Reference sample :
  - 1_1: ATATGTACGTAGTA
  - 1_2: ATATCTACGAAGTA
  - 1_3: ATATCTACTAAGTA

  } Species 1

  - 2_1: ATATGTACGTAGTT
  - 2_2: ATATGTACGAAGTT
  - 2_3: ATTTCTACTAAGTT

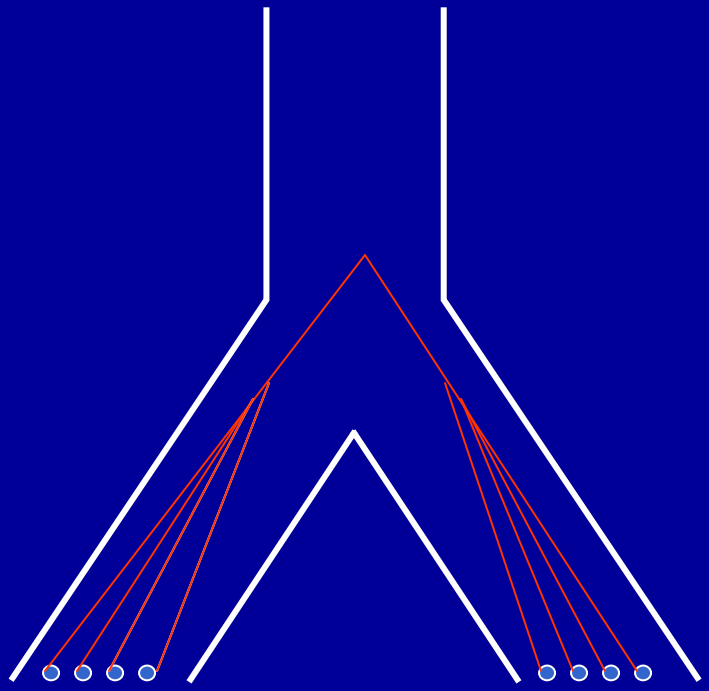
  } Species 2

- Individual to be tested:
  - X_1 : ATATGTACCTAGTA
  - X_2 : TTATCTACCTAGAA

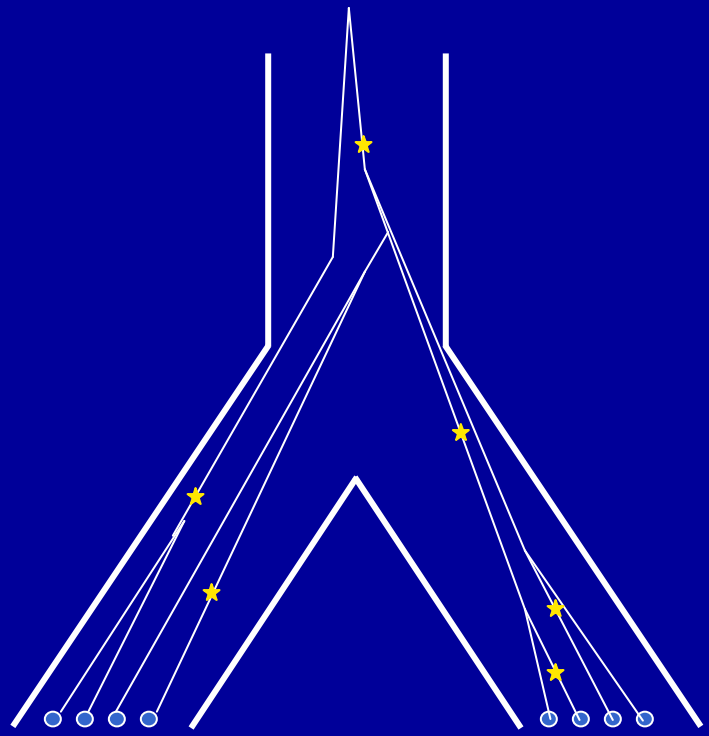
Questions

- What is the best assignation methods?
- What sample size is needed?
- What level of polymorphism is needed?

Two species can be separated but the coalescent trees can still be mixed

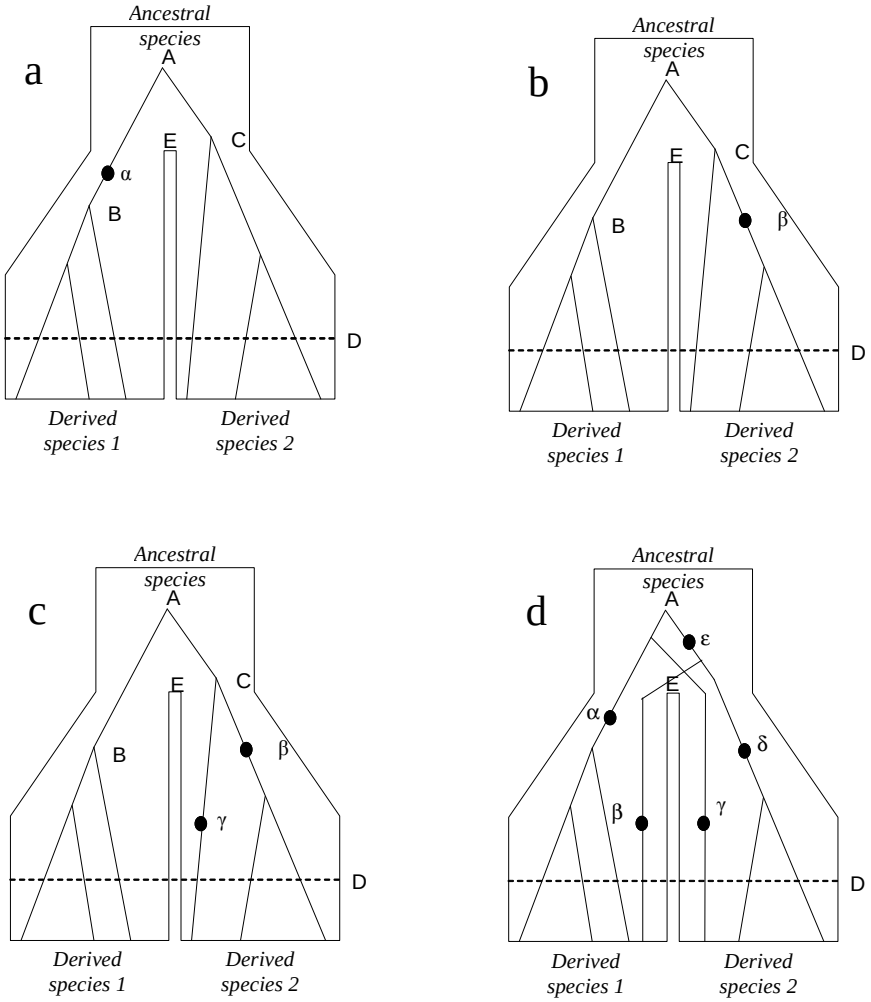


- Speciation gene

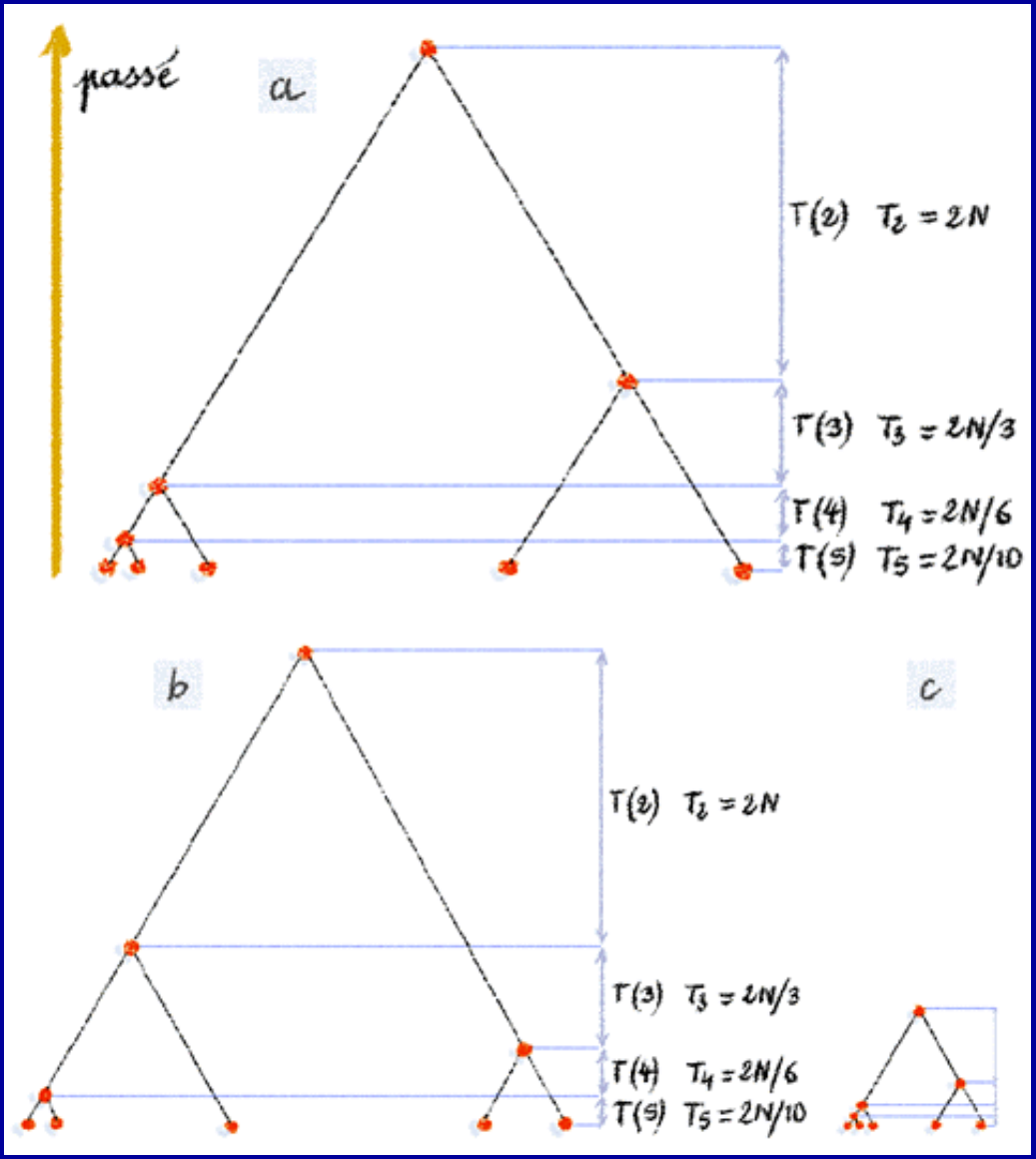


- Neutral gene

Several cases

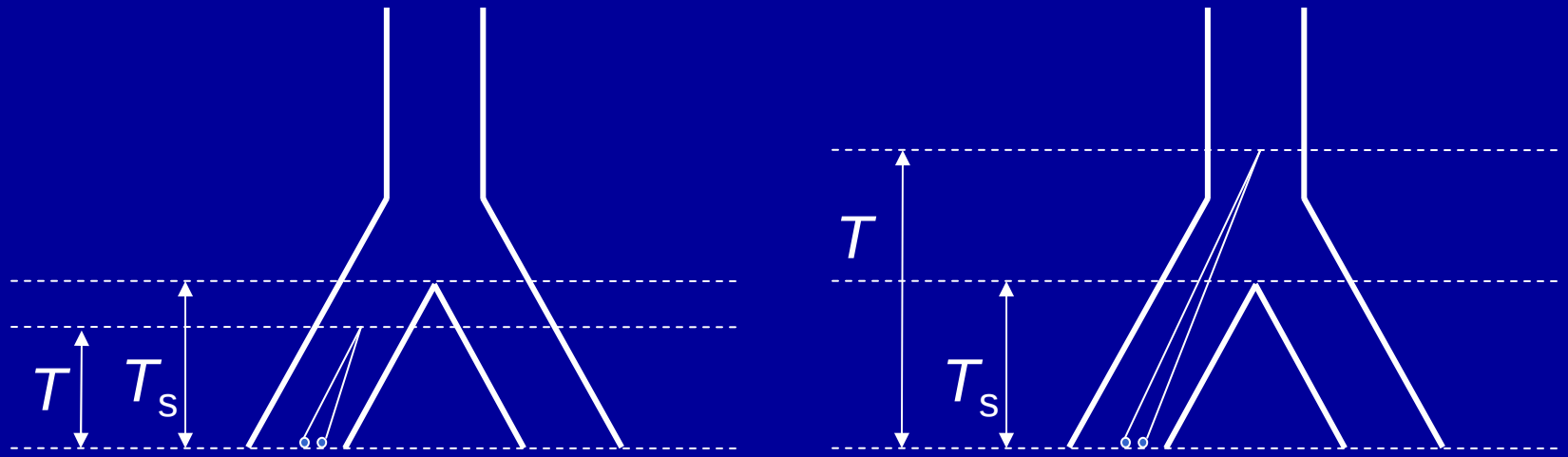


Arbre de coalescence moyens



(source L. Excoffier, La Recherche)

Cas de deux espèces



locus haploïde à transmission maternelle : $P(T > Ts) = (1-1/N_f)^T = e^{-T/N_f}$

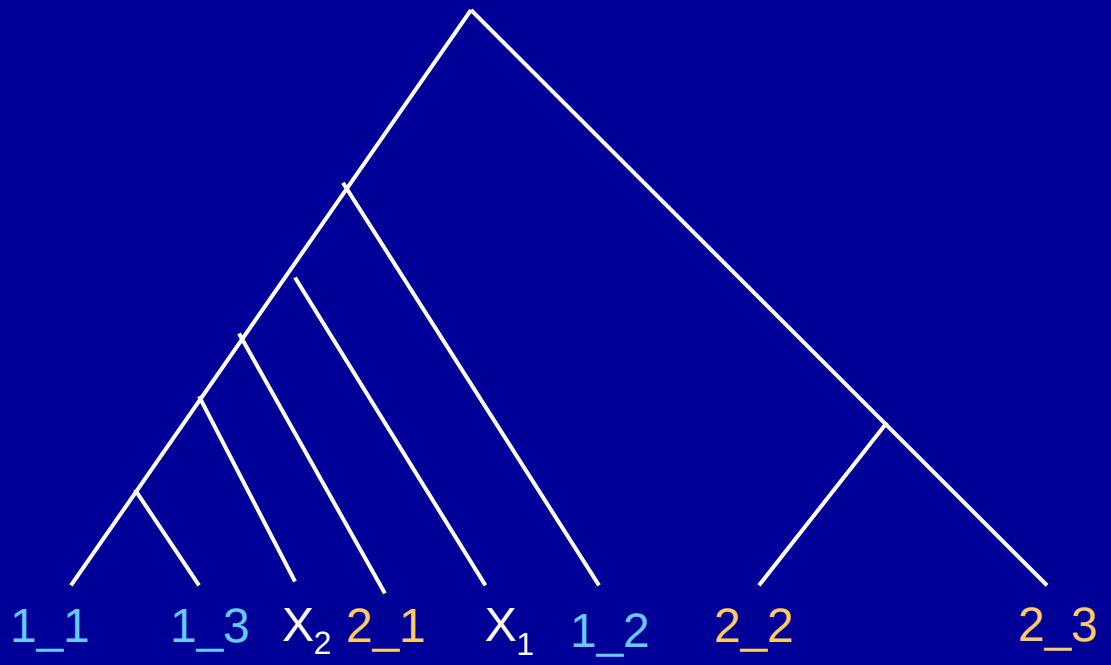
locus nucléaire à transmission biparentale : $P(T > Ts) = (1-1/2N)^T = e^{-T/2N}$

si sexe ratio équilibré, $N_f = N_m = N/2$, pour des espèces séparées depuis N générations,

$P(T>N)=e^{-2} \approx 0.135$ (locus haploïde)

$P(T>N)=e^{-1/2} \approx 0.607$ (locus diploïde)

Phylogenetic methods



- Phylogenetic method used: maximum likelihood (PhyML, Guindon and Gascuel 2003) or neighbor joining.

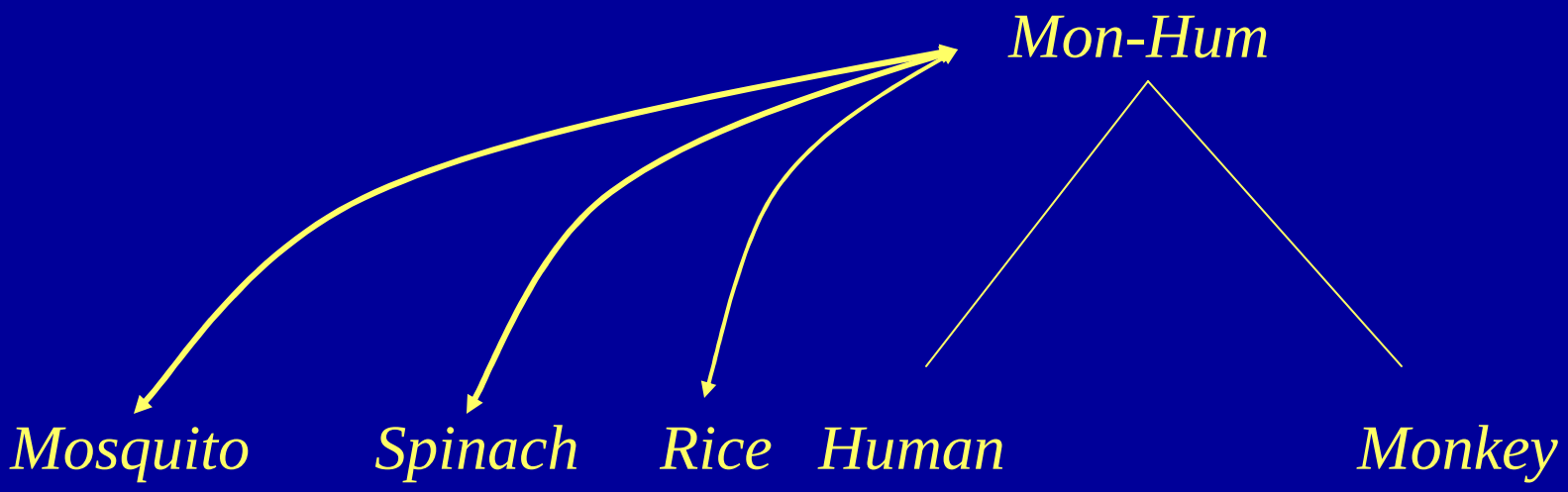
- Majority rule: X_1 and X_2 classified as belonging to species 1, because they cluster in majority with individuals of species 1.

Distance Matrix

PAM	Spinach	Rice	Mosquito	Monkey	Human
Spinach	0.0	84.9	105.6	90.8	86.3
Rice	84.9	0.0	117.8	122.4	122.6
Mosquito	105.6	117.8	0.0	84.7	80.8
Monkey	90.8	122.4	84.7	0.0	**3.3**
Human	86.3	122.6	80.8	**3.3**	0.0

First Step

PAM distance 3.3 (Human - Monkey) is the minimum. So we'll join Human and Monkey to MonHum and we'll calculate the new distances.



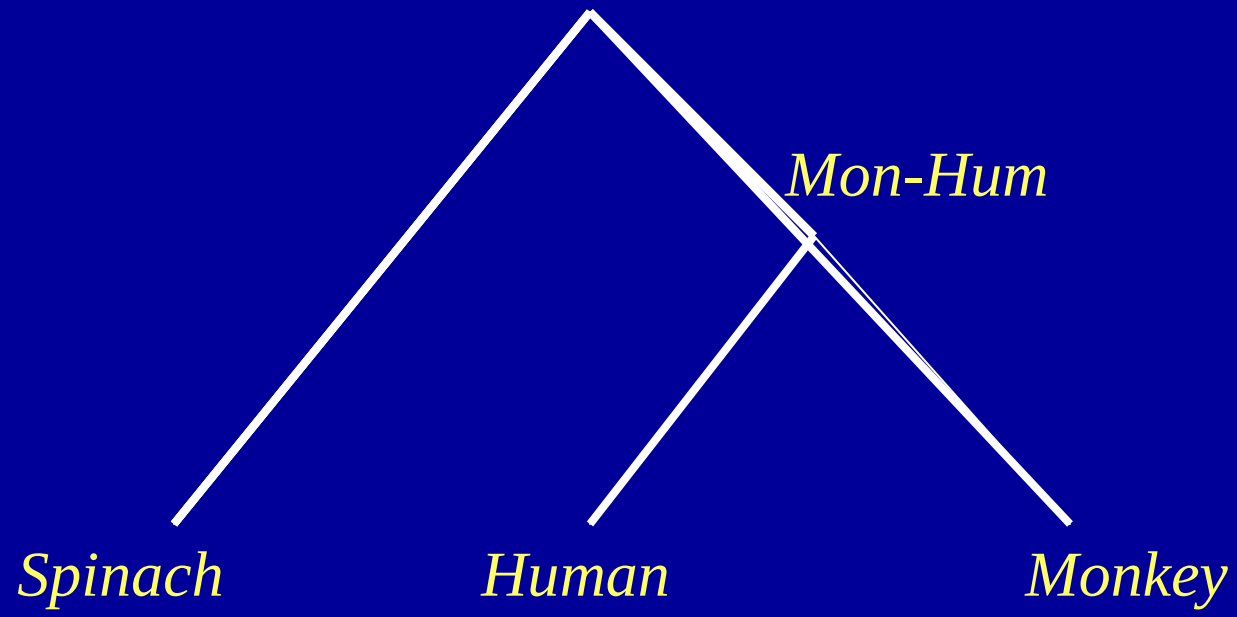
Calculation of New Distances

After we have joined two species in a subtree we have to compute the distances from every other node to the new subtree. We do this with a simple average of distances:

Dist[Spinach, MonHum]

= (Dist[Spinach, Monkey] + Dist[Spinach, Human])/2

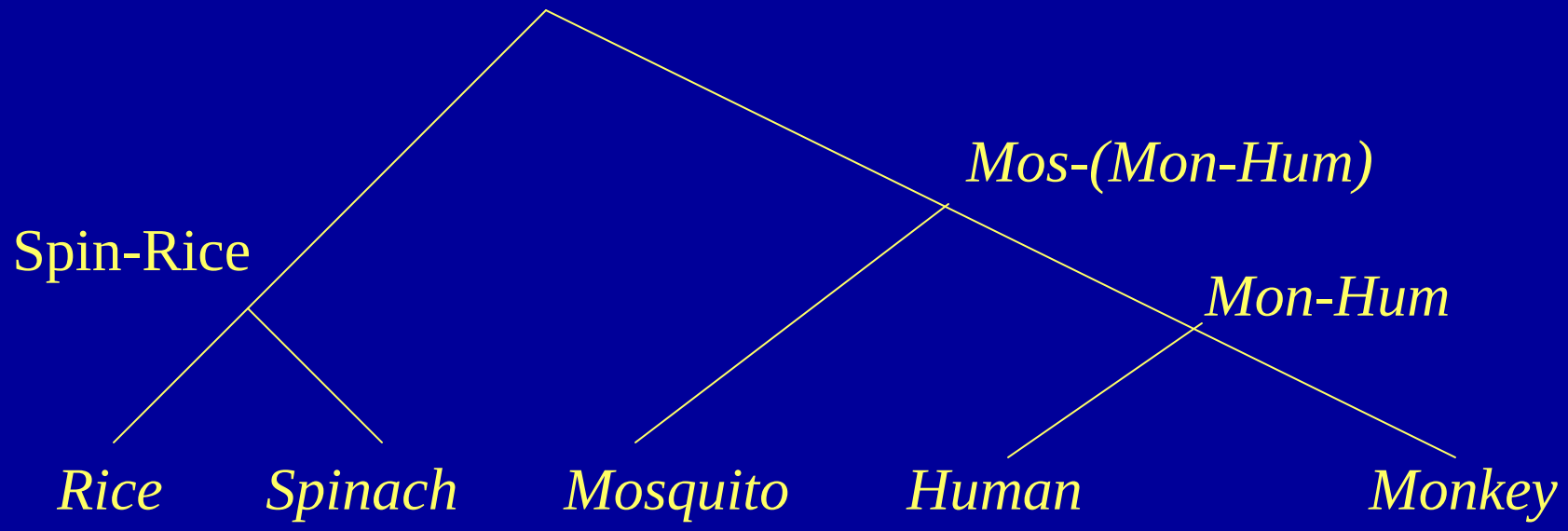
= (90.8 + 86.3)/2 = 88.55



Last Joining

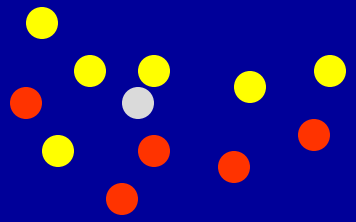
PAM	SpinRice	MosMonHum
Spinach	0.0	108.7
MosMonHum	108.7	0.0

(Spin-Rice)-(Mos-(Mon-Hum))



k = 1 nearest neighbor classifier

- Compute the Kimura 2 Parameter (K2P) distance between the test individual and all individuals of the reference sample.

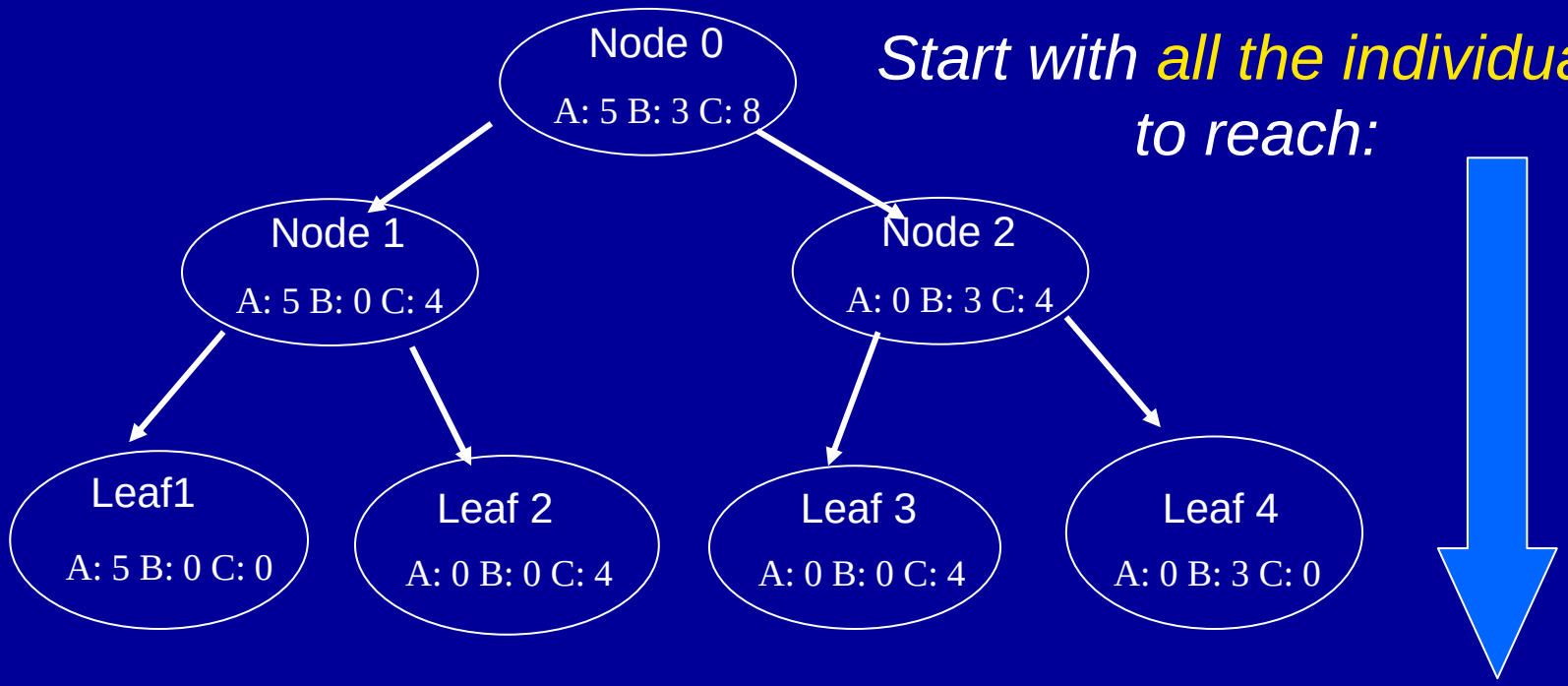


- The test individual is classified in the species of its nearest neighbor.

CART (Classification And Regression Tree)

(Breiman et al., 1984, 1996)

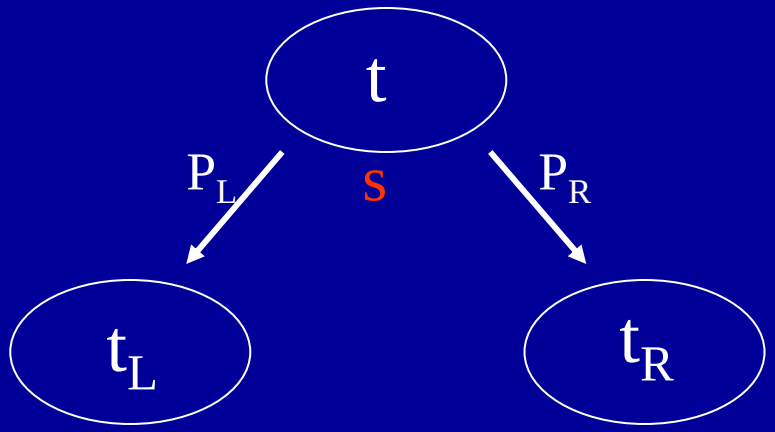
- Builds a classification tree from the reference sample



Start with all the individuals to reach:

« individuals in each leaf belong to the same species »

Finding the best split



t: set of individuals

S: set of variables

$s \in S$

$P_L + P_R = 1$

Decrease in impurity: $\Delta I(s,t) = I(t) - p_L\, I(t_L) - p_R\, I(t_R)$

Rule to select a splitting candidate:

s^* selected as $\Delta I(s^*,t) = \max \{ \Delta I(s,t), s \in S \}$

Stop splitting rule: e.g. threshold β

$\max \{ \Delta I(s,t), s \in S \} < \beta$

Computing the impurity of the nodes

node t = subset of individuals

p (j | t) = relative proportion of individuals of class *j* in node t

Impurity criterion at node t: $I(t) \doteq \Phi\left(p(1|t),\ldots, p(j|t),\ldots, p(J|t)\right)$

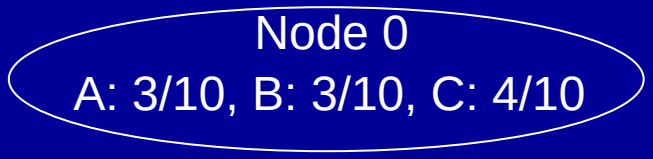
$I(t) = -\sum_j p(j|t) \log p(j|t)$ ⟵ *entropy*

$I(t) = 1 - \sum_j p(j|t)^2$ ⟵ *Gini index*

$$\begin{cases} \Phi\left(\frac{1}{J},\ldots,\frac{1}{J}\right) \longrightarrow \text{maximum} \\ \Phi \quad \text{minimum for} \quad (1,0,\ldots,0),\ldots,(0,\ldots,0,1) \end{cases}$$

Example: 3 species, 10 individuals, 4 variables

Species	x1	x2	x3	x4
A	a	g	c	g
A	a	g	c	g
A	a	a	a	g
B	t	t	a	g
B	t	t	a	g
B	t	a	c	g
C	a	c	a	g
C	g	c	a	g
C	a	c	a	g
C	a	c	c	g



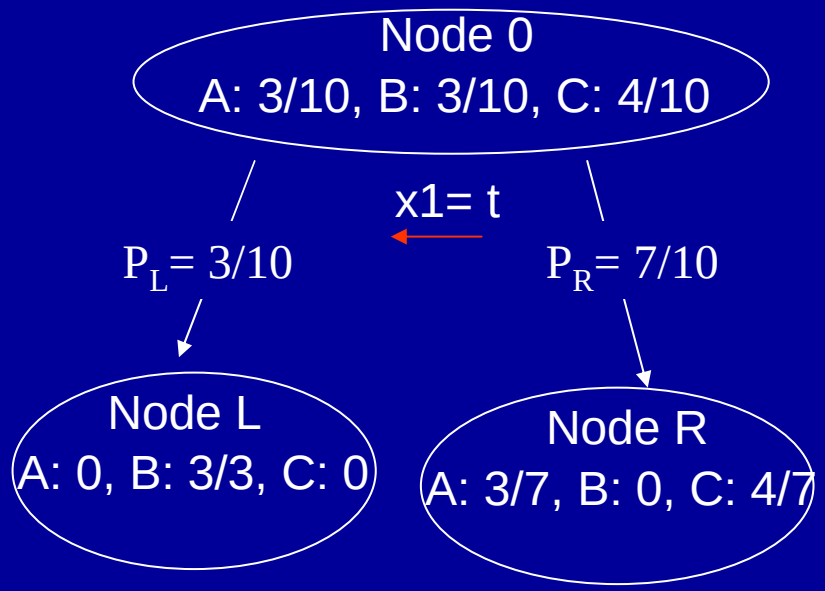
At node 0: $I(t) = I(node0) = - \sum_j p(j|t) \log p(j|t)$

$I(node0) = - [3/10 \times \log(3/10) + 3/10 \times \log(3/10) + 4/10 \times \log(4/10)]$

$I(node0) = 1.0889$

Splitting according to x_1

Species	x1	x2	x3	x4
A	a	g	c	g
A	a	g	c	g
A	a	a	a	g
B	t	t	a	g
B	t	t	a	g
B	t	a	c	g
C	a	c	a	g
C	g	c	a	g
C	a	c	a	g
C	a	c	c	g



$I(t) = - \sum_j p(j|t) \log p(j|t)$

At node L: $I_{x1}(\text{nodeL}) = - [0 + 3/3 \times \log(3/3) + 0] = 0$

At node R: $I_{x1}(\text{nodeR}) = - [3/7 \times \log(3/7) + 0 + 4/7 \times \log(4/7)]$

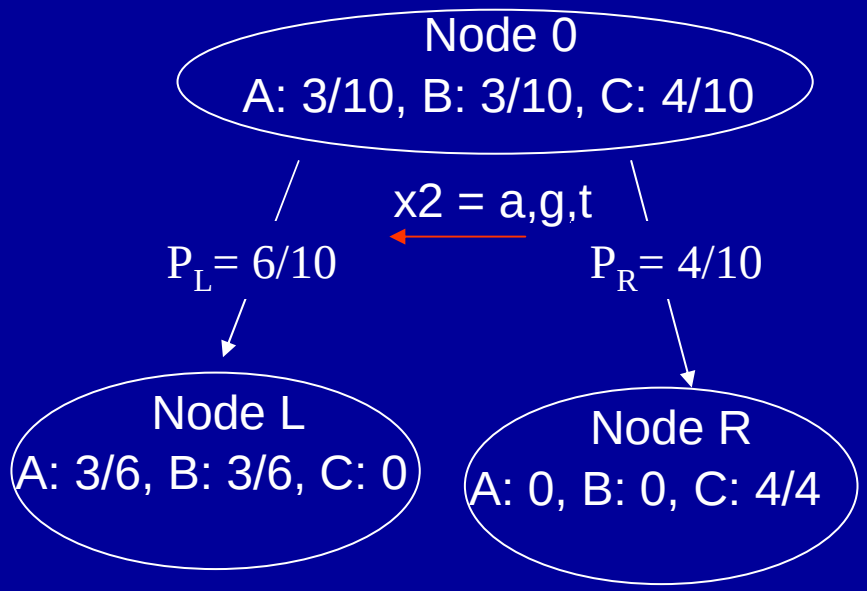
$I_{x1}(\text{nodeR}) = 0.6829$

$\Delta I(x_1,t) = I(\text{node0}) - P_L * I_{x1}(\text{nodeL}) - P_R * I_{x1}(\text{nodeR})$

$\Delta I(x_1,t) = 1.0889 - 0.3 \times 0 - 0.7 \times 0.6829 = 0.6109$

Splitting according to x_2

Species	x1	x2	x3	x4
A	a	g	c	g
A	a	g	c	g
A	a	a	a	g
B	t	t	a	g
B	t	t	a	g
B	t	a	c	g
C	a	c	a	g
C	g	c	a	g
C	a	c	a	g
C	a	c	c	g



$$I(t) = -\sum_j p(j|t) \log p(j|t)$$

At node L: $I_{x2}(\text{nodeL}) = -[3/6*\log(3/6) + 3/6*\log(3/6) + 0]$

$I_{x2}(\text{nodeL}) = 0.6931$

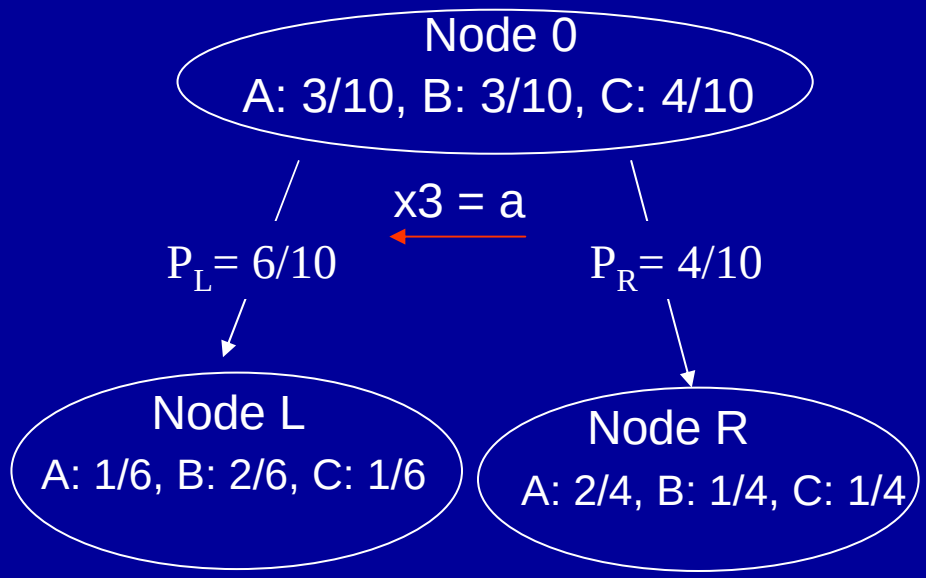
At node R: $I_{x2}(\text{nodeR}) = -[0 + 0 + 4/4*\log(4/4)] = 0$

$$\Delta I(x_2,t) = I(\text{node0}) - P_L * I_{x2}(\text{nodeL}) - P_R * I_{x2}(\text{nodeR})$$

$$\Delta I(x_2,t) = 1.0889 - 0.6 * 0.6931 - 0.4 * 0 = 0.6730$$

Splitting according to x_3

Species	x1	x2	x3	x4
A	a	g	c	g
A	a	g	c	g
A	a	a	a	g
B	t	t	a	g
B	t	t	a	g
B	t	a	c	g
C	a	c	a	g
C	g	c	a	g
C	a	c	a	g
C	a	c	c	g



$$I(t) = -\sum_j p(j|t) \log p(j|t)$$

At node L: $I_{x3}(\text{nodeL}) = -[1/6*\log(1/6) + 2/6*\log(2/6) + 2/6*\log(2/6)] = 1.031$

At node R: $I_{x3}(\text{nodeR}) = -[2/4*\log(2/4) + 1/4*\log(1/4) + 1/4*\log(1/4)] = 1.040$

$$\Delta I(x_3,t) = I(\text{node0}) - P_L * I_{x2}(\text{nodeL}) - P_R * I_{x2}(\text{nodeR})$$

$$\Delta I(x_3,t) = 1.0889 - 0.6 * 1.031 - 0.4 * 1.040 = 0.0662$$

Choosing the best split

Species	x1	x2	x3	x4
A	a	g	c	g
A	a	g	c	g
A	a	a	a	g
B	t	t	a	g
B	t	t	a	g
B	t	a	c	g
C	a	c	a	g
C	g	c	a	g
C	a	c	a	g
C	a	c	c	g

$\Delta I(x_1,t) = 0.6109$

$\Delta I(x_2,t) = 0.6730$

$\Delta I(x_3,t) = 0.0662$

$x_4 \longrightarrow$ no division

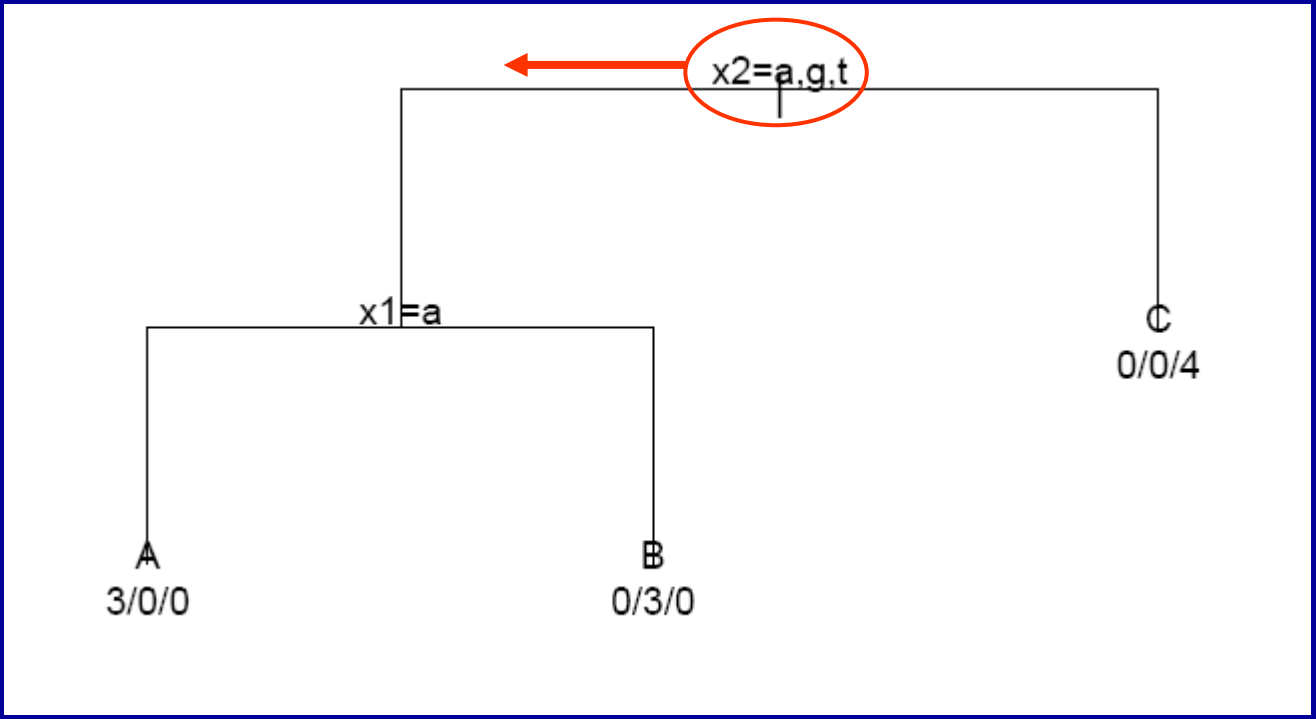
$$\Delta I(s^*,t) = \max_{s \in S} \Delta I(s,t)$$

x_2 is selected

Implementation

Software: R
Package: rpart

Criterion: *Gini*



Random Forest

- Draw *n* bootstrap samples from the reference sample.

- For each bootstrap samples, grow a classification tree, with *m* randomly sampled sites and choose the best tree from these variables.

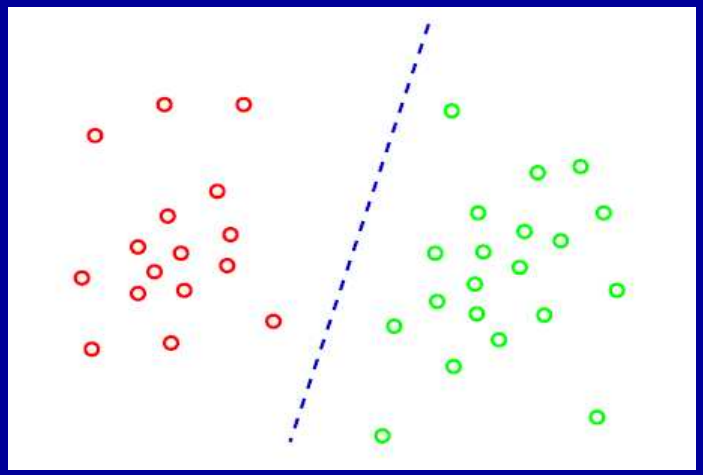
- Predict the species of the test individual by aggregating the predictions of the *n* trees (majority vote).

↳This method is more robust than CART because it reduces the effect of the first split.

Breiman (2001). Random Forests. Machine Learning 45 (1), 5-32.

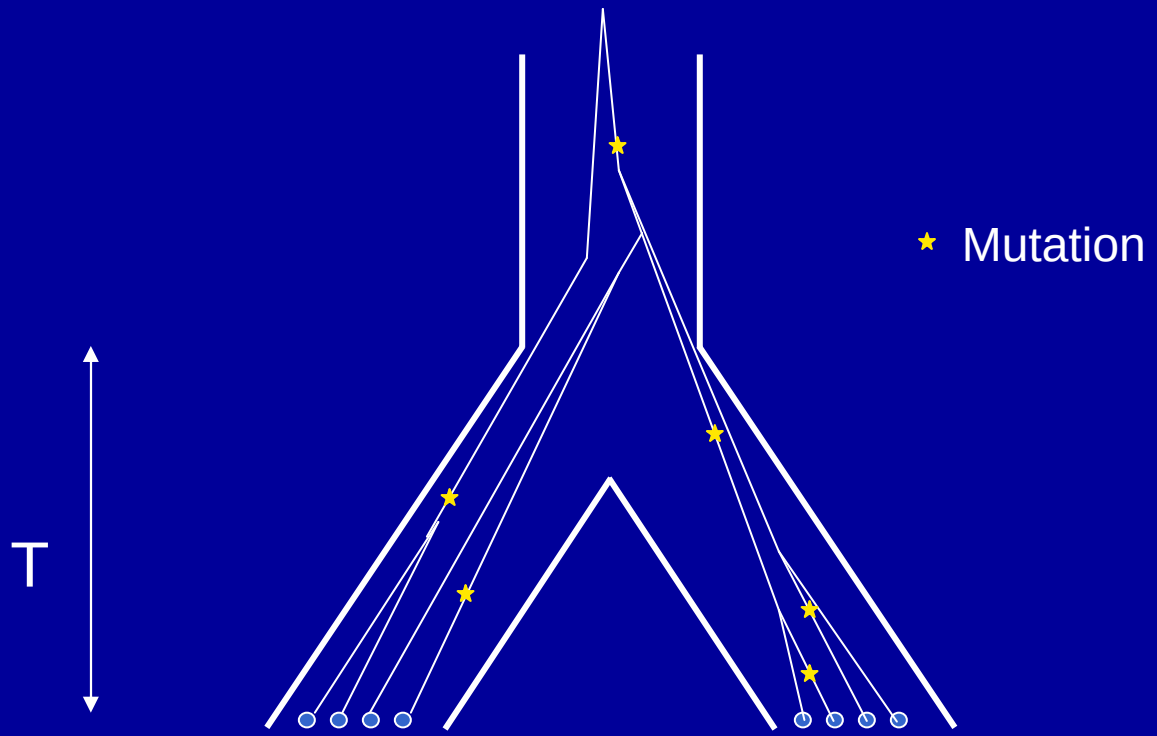
Kernel methods

- Project the data into a space of high dimension.
- Find a hyperplane that best separates the two species in this high-dimensional space.
- The species of a new individual is predicted by checking what side of the hyperplane it is on.



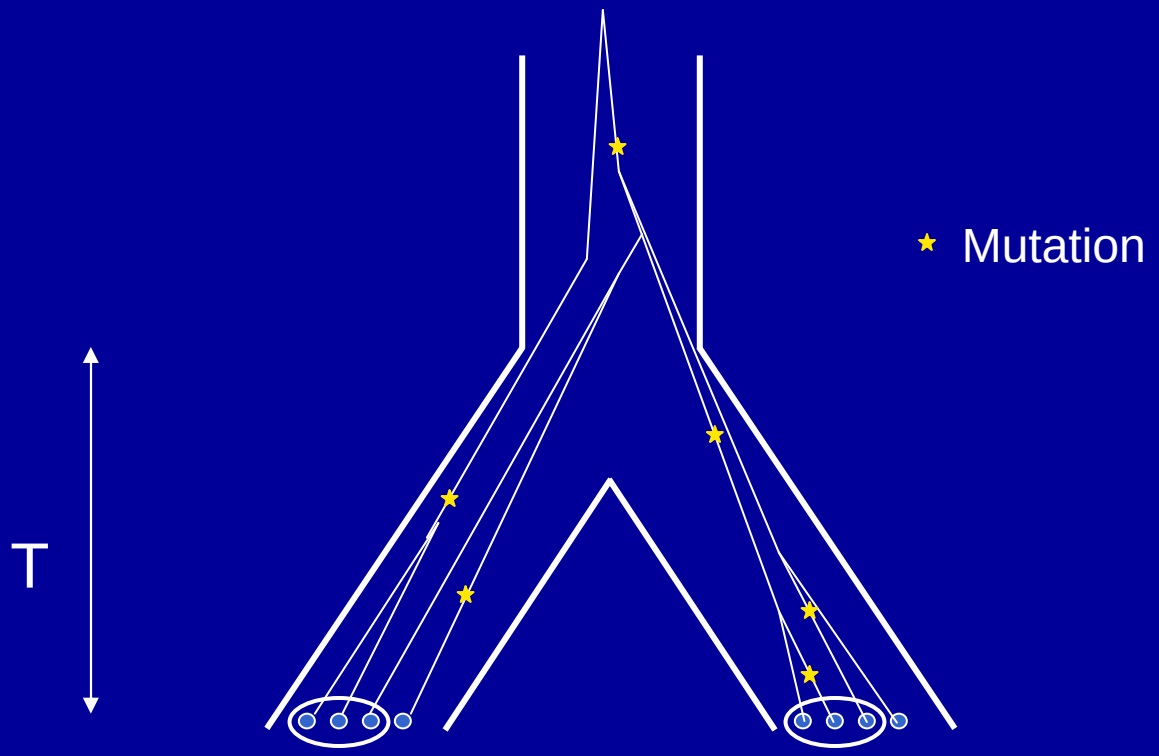
Cortes C., Vapnik V. *Mach Learn* 1995, **20**:273-297

Simulation method



- Simulations performed with simcoal 2 (Laval & Excoffier, 2004, *BioInformatics*)
- one ancestral species that splits into two (or more) species T generations ago in the past.
- The ancestral species and the two new species are of constant size, with N_f females.
- Sequences with mutation rate μ → parameter of interest $\theta = 2N_f\mu$

Evaluation of the different classification methods



- We simulate *n* +1 individuals in each species.
- *n* individuals in each population are considered as the reference sample, and the last one as the individual to test.
- Using repeated simulations, we compute the proportion of cases in which each test individual is correctly assigned to its species.

Parameters assumed for the simulation study

- θ = 3, 12 or 30
- Reference sample size n = 3, 5, 10, 25
- Effective female population size: N_f = 1000
- Separation time T = 100, 500, 1000, 5000 or 10000.
- Number of species: n_s = 2 to 5.

Effect of speciation time on the success rate of the different method for low polymorphism

(2 species, Reference sample size = 10, $\theta = 3$)

Speciation time	NJ	PhyML	1-NN	CART	RF	Kernel
100	62.90%	62.25%	**65.45%**	65.40%	64.30%	64.95%
500	**87.25%**	86.30%	87.20%	87.15%	86.40%	87.15%
1000	95.90%	96.00%	**96.75%**	96.55%	96.00%	95.75%
5000	**100%**	**100%**	**100%**	99.85%	**100%**	99.80%
10000	**100%**	**100%**	**100%**	**100%**	**100%**	99.90%

- Increase of success rate with separation time
- Best performance of 1-NN and NJ

Effect of speciation time on the success rate of the different method for high polymorphism

(2 species, Reference sample size = 10, $\theta = 30$)

Speciation time	NJ	PhyML	1-NN	CART	RF	Kernel
100	75.60%	75.30%	76.25%	75.50%	**77.75%**	73.45%
500	96.10%	**96.20%**	95.55%	93.50%	95.25%	94.00%
1000	**99.15%**	**99.15%**	98.55%	97.10%	98.35%	96.90%
5000	99.95%	**100%**	**100%**	99.40%	**100%**	99.45%
10000	**100%**	**100%**	**100%**	99.40%	**100%**	99.55%

- Increase of success rate with separation time
- Best performance of 1-NN, PhyML and RF

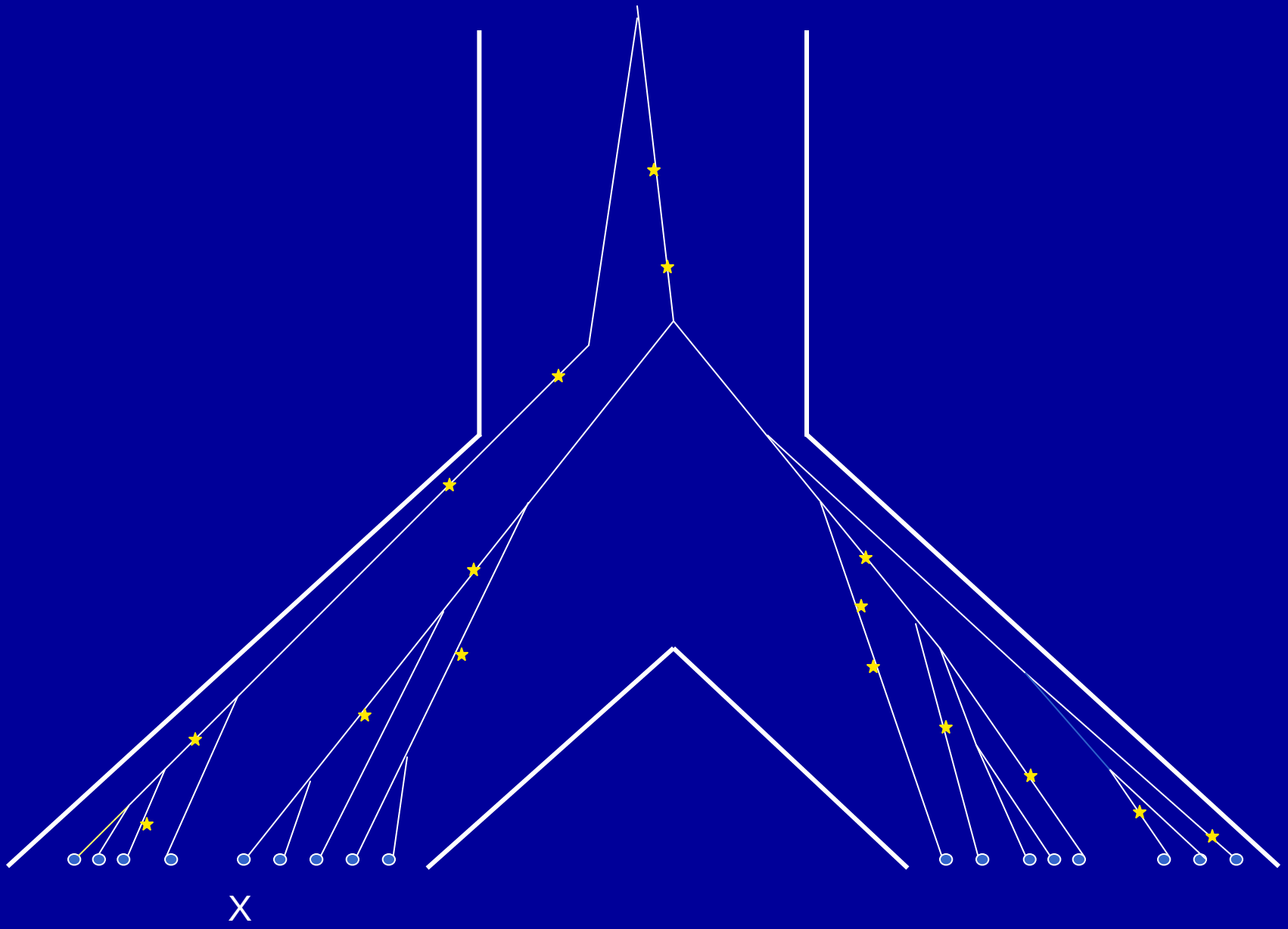
Effect of the size of the reference sample

(Separation time = 500, θ = 3)

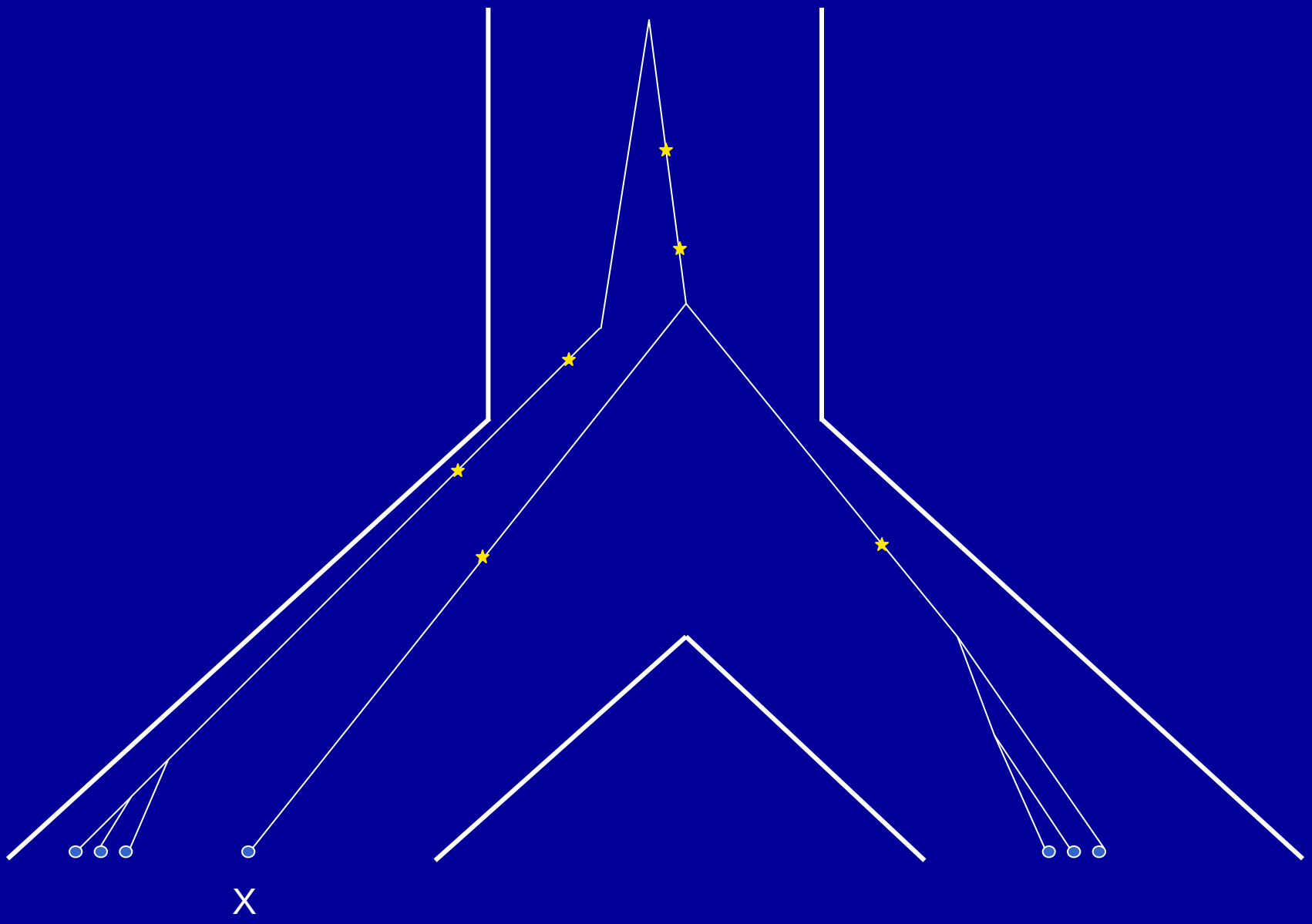
Reference sample size	NJ	PhyML	1-NN	CART	RF	Kernel
3	77.45%	77.50%	**78.05%**	77.15%	77.35%	76.15%
5	**84.20%**	83.85%	83.30%	82.95%	82.40%	82.10%
10	**87.25%**	86.30%	87.20%	87.15%	86.40%	87.15%
25	**92.00%**	91.70%	91.10%	90.80%	89.40%	90.75%

- Strong increase of performance with sample size
- Best performance of 1-NN and NJ

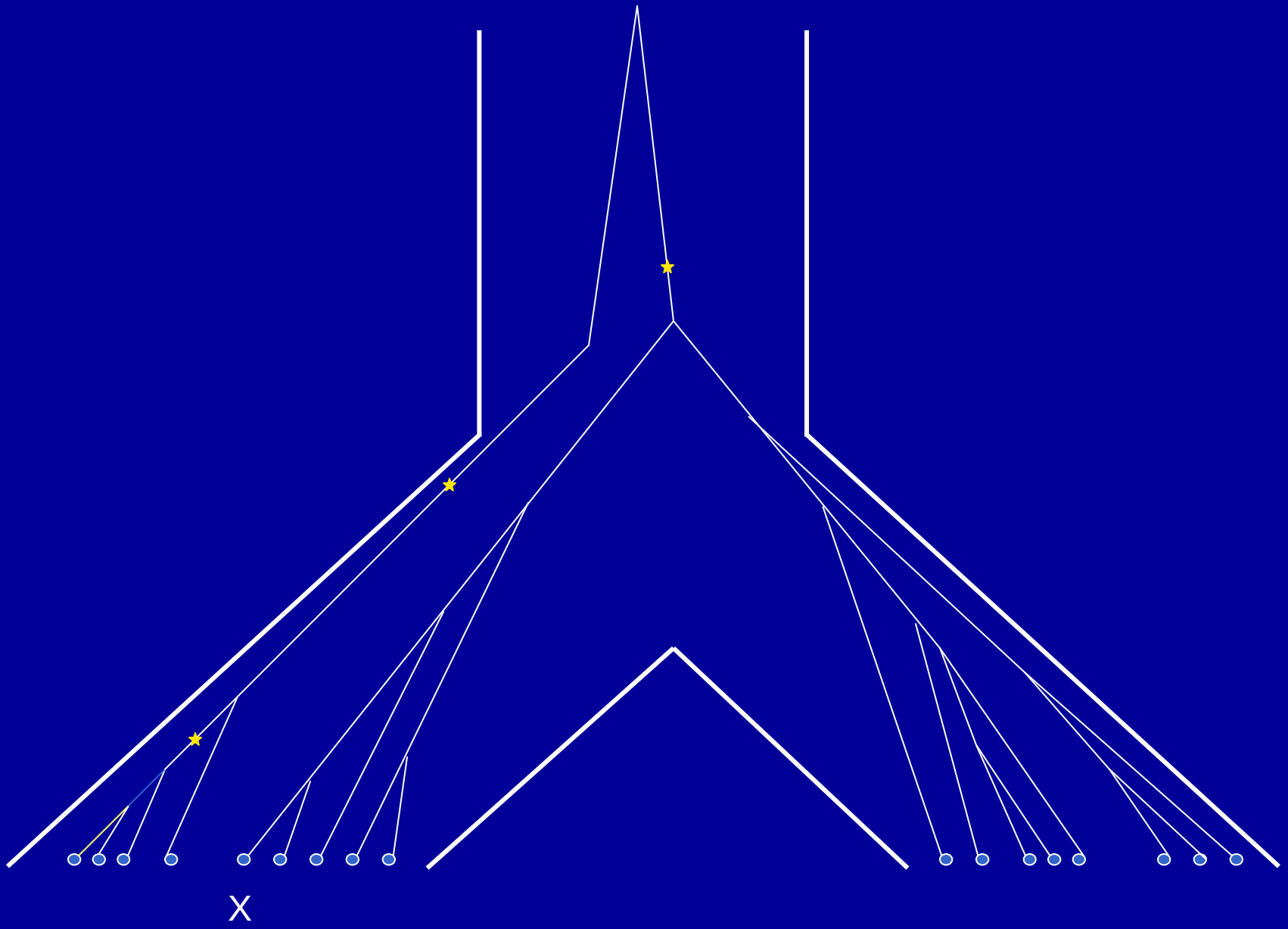
Large sample size



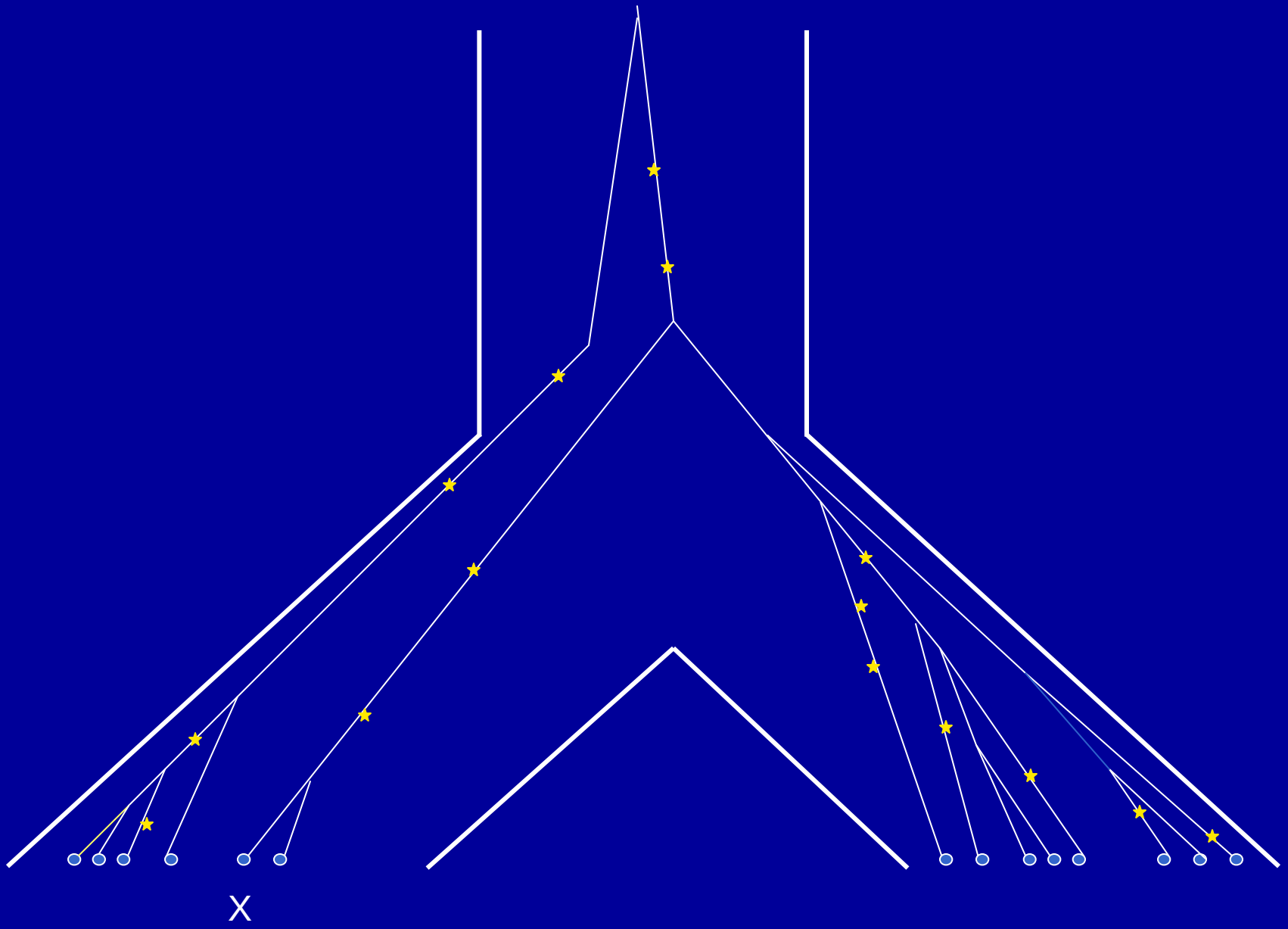
Small sample size



Lower mutation rate



The efficiency of NJ and 1-NN may stem from their focus on the closest neighbor



Effect of the number of species for low polymorphism

Number of species	NJ	PhyML	1-NN	CART	RF	Kernel
2	87.25%	86.30%	**87.30%**	87.15%	87.20%	87.15%
3	**81.73%**	80.77%	80.67%	80.40%	80.97%	81.10%
4	75.80%	75.00%	75.40%	75.68%	**75.95%**	74.78 %
5	**73.26%**	72.36%	72.58%	72.84%	73.22%	70.74%

(Separation time = 500, $\theta = 3$)

Effect of the number of species for high polymorphism

Number of species	NJ	PhyML	1-NN	CART	RF	Kernel
2	96.10%	**96.20%**	95.55%	93.50%	95.25%	94.00%
3	**94.40%**	94.23%	94.00%	90.93%	93.50%	92.10%
4	**93.78%**	93.73%	92.90%	90.10%	92.53%	91.40%
5	92.46%	92.38%	**92.70%**	88.98%	92.08%	90.46%

(Separation time = 500, θ = 30)

Adding nuclear loci

- We considered the case where data for nuclear loci are also available.
- We assumed that
  - these loci are independent.
  - they all have the same θ (= $4N\mu$) value, equal to the value for the cytoplasmic genes.

Combining the different loci

- The performance of each locus for a given method is assessed using the reference sample.
- Then the unknown individuals are assigned to a given species by making a vote among the loci
  - the weight of each locus being its capacity to correctly assign the individuals.

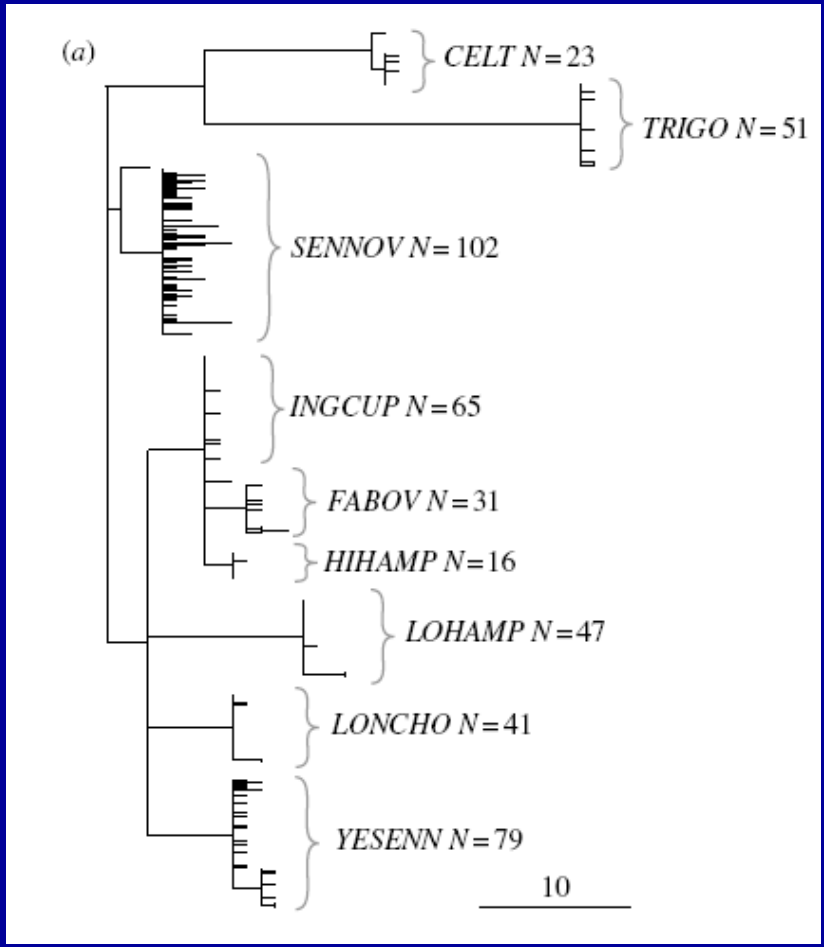
Adding nuclear genes

Number of nuclear loci	NJ	PhyML	1-NN	CART	RF	Kernel
0	87.25%	86.30%	**87.30%**	87.15%	86.40%	87.15%
1	88.05%	87.10%	**91.40%**	89.70%	83.55%	83.70%
2	90.60%	90.35%	**95.00%**	93.20%	86.25%	86.80%
3	92.80%	92.40%	**96.20%**	95.05%	88.95%	89.75%
4	94.70%	94.60%	**97.70%**	96.55%	91.30%	91.90%

- 2 populations
- θ = 3, separation time = 500, reference sample size = 10

Astraptes fulgeraptor

Fig. 1. Newly eclosed female *A. fulgerator* (species LOHAMP, voucher code 02-SRNP-9770) from the ACG.



- Good level of separation
- Data on the barcoding sequence (part of CO1)

Hebert et al 2004. Proc. Natl Acad. Sci. USA 101,14812–14817

The cowries

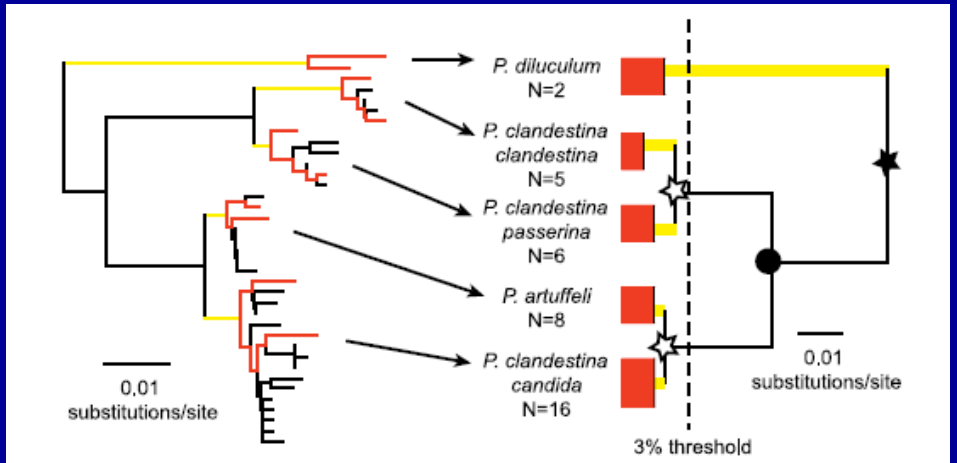


Figure 3. Intraspecific and Interspecific Estimations

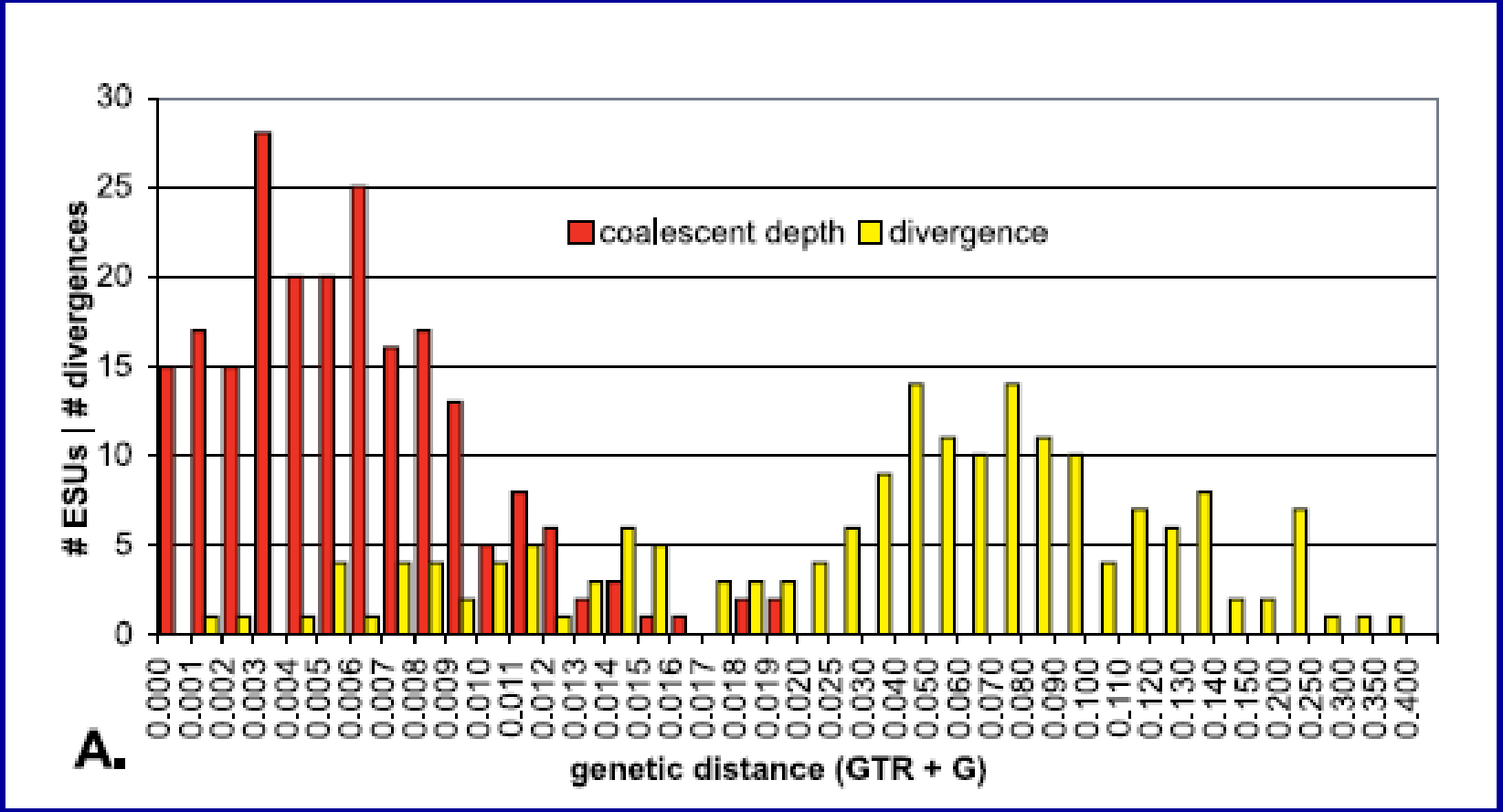
A subclade of five cowrie ESUs shows how both coalescent and divergence depths are generated. The two most disparate individuals are culled from within each ESU (left—red) and used in a constrained phylogeny with a molecular clock enforced (right) to recover both the maximum coalescent depth (red) and the divergence depths between sisters (yellow). Two young ESUs (stars) would be missed (false negatives) if a 3% threshold cutoff (shown) was employed. Note that *Palmadusta artuffeli,* a Japanese endemic species, is nested among monophyletic subspecies of the paraphyletic species *P. clandestina*. The black circle indicates the coalescent for the species *P. clandestina,* and the black star indicates the interspecific divergence for species-level analyses.

DOI: 10.1371/journal.pbio.0030422.g003

- "Species" much less separated, low polymorphism
- Data on the barcoding sequence (part of CO1)

Meyer et al (2005) PLoS Biol 3: e422

Overlap between coalescent dept and divergence



Neotropical butterflies: Subfamily Ithomiinae

Hypothyris euclea

- Data on the barcoding sequence (part of CO1), on a larger part of CO1 and on a nuclear locus (EF1a)

Elias et al (2007) Limited performance of DNA barcoding in a diverse community of tropical butterflies. Proc R Soc B 274:2881-2889

Results on experimental data

Data set	NJ	PhyML	1-NN	CART	RF	Kernel
Astraptes	99.36%	99.36%	99.36%	98.22%	99.36%	**99.57%**
Cowries (species)	**95.45%**	93.40%	**95.45%**	78.40%	94.65%	94.45%
Cowries (subspecies)	91.10%	86.37%	91.31%	72.38%	**91.41%**	89.20%
Amazonian Butterflies (barcode)	91.73%	91.20%	90.40%	75.47%	92.00%	**92.80%**
Amazonian Butterflies (mtDNA)	91.47%	91.20%	91.73%	71.20%	92.00%	**93.60%**
Amazonian Butterflies (nuclear gene)	87.74%	**90.32%**	**90.32%**	52.90%	80.64%	89.03%

- Differences in success of the method depending on the data set.

Conclusions

- Most methods perform correctly, but the level of polymorphism and sample sizes are important factors.
- Some methods (1-NN, NJ) generally perform slightly better, but the best strategy might be to test on your own reference sample.
- Adding nuclear loci does increase the performance.
- More details in the paper.

Austerlitz, F., David, O., Schaeffer, B., Bleakley, K., Olteanu, M., Leblois, R., Veuille, M. & Laredo, C. 2009 DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC bioinformatics* **10 (Suppl 14)**, S10.