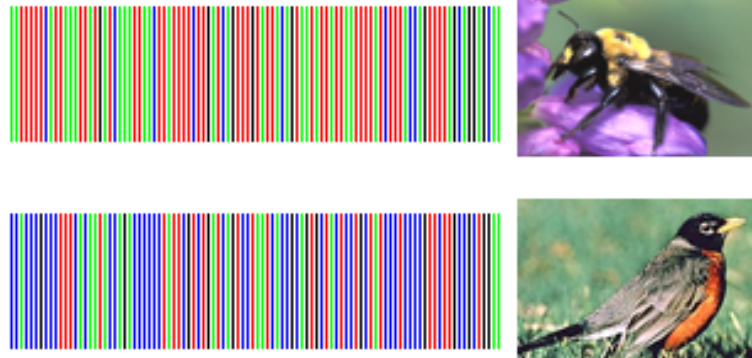


Problèmes et limites de l'approche "code barre ADN"/"DNA Barcode"



Raphaël Leblois, MC MNHN, dep^t Systématique & Evolution
Origine, Structure et Evolution de la biodiversité
(UMR CNRS/MNHN/IRD 5202)

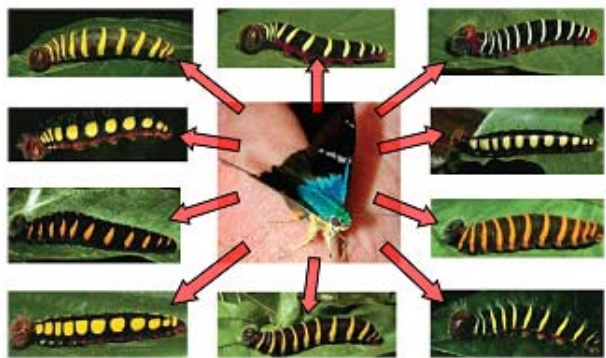


Plan de l'exposé

1. Les grandes lignes du projet "code barre ADN"
 - Qu'est ce que c'est?
 - Bénéfices, Problèmes et Limites potentiels
 - Premiers outils d'analyse
2. Arbres d'espèces et arbres de gènes
3. Comment évaluer et améliorer le projet "code barre ADN"



1. Le projet "code barre ADN" Qu'est ce que c'est?



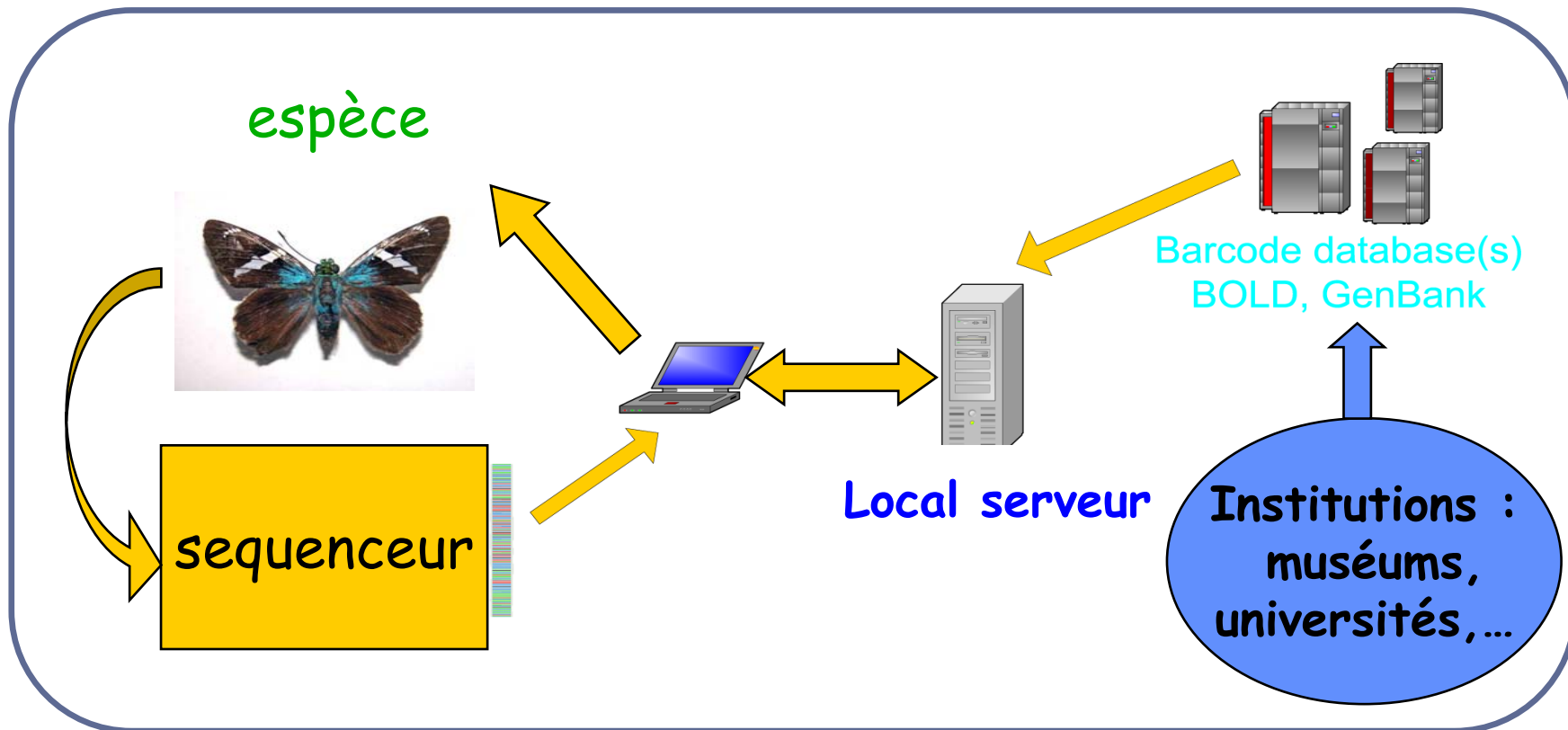
?

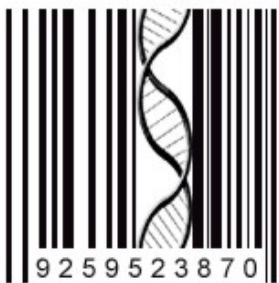
Approche standardisée
d'identification des espèces
de plantes et d'animaux à
partir d'une courte séquence
d'ADN (appelée code barre ADN)

= **technique/outils** d'identification taxonomique
utilisable par des non spécialistes



1. Le projet "code barre ADN" Qu'est ce que c'est?





1. Le projet "code barre ADN" Qu'est ce que c'est?

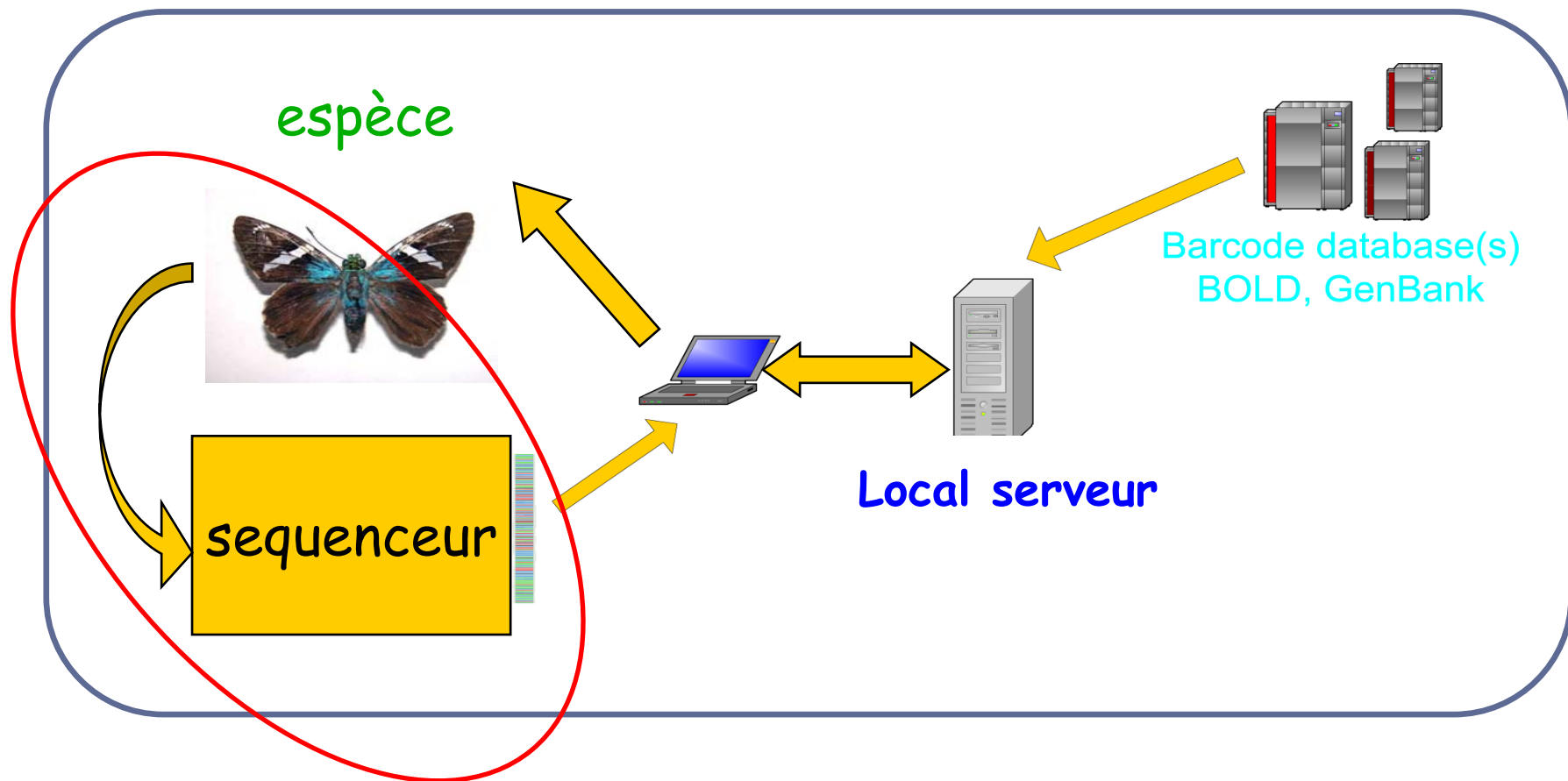
CONSORTIUM FOR THE BARCODE OF LIFE

<http://barcoding.si.edu/>

- Collaboration de muséums/ herbiers/zoos/jardins botaniques avec des scientifiques en taxonomie, génétique des populations, informatique, des bioo-techs,... (100 institutions dans 40 pays).
- **But** : -promouvoir le "code barre AND" (données, analyse)
 - Développer une base de données publique
 - Encourager le développement d'outils d'identification ("barcoder")

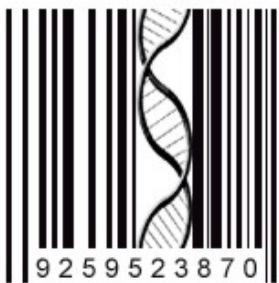


1. Le projet "code barre ADN" Qu'est ce que c'est?



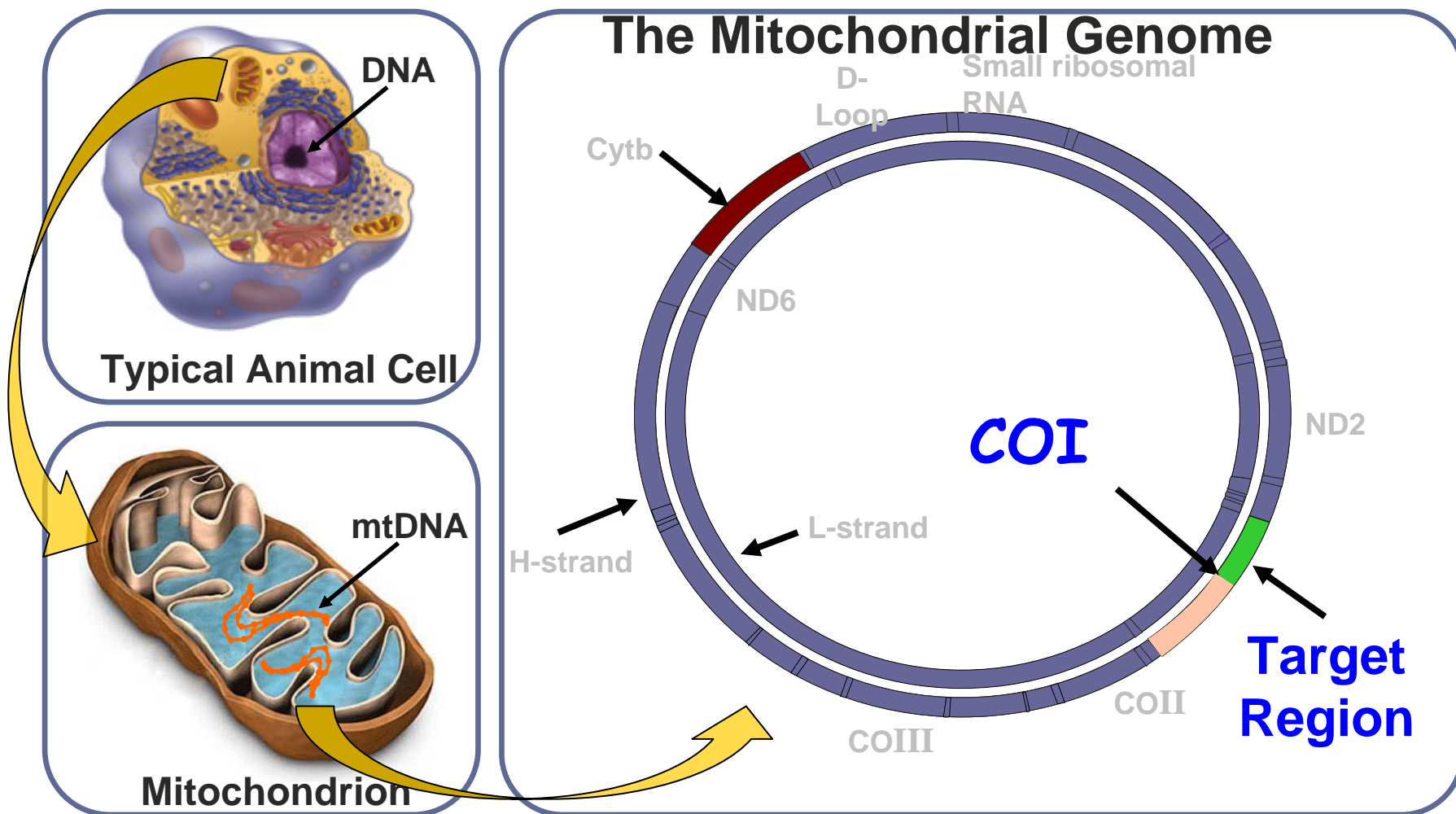
Choix des séquences ADN "code barre"

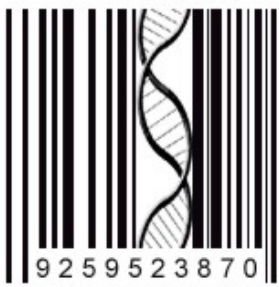
Protocoles d'extraction et de séquencage



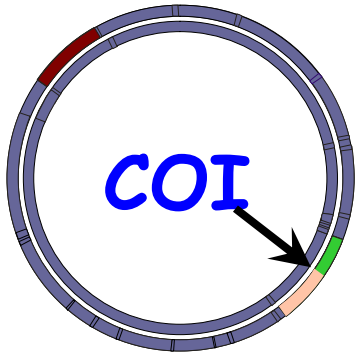
1. Le projet "code barre ADN" quel séquence d'ADN?

650pb du gène *COI* (cytochrome C oxidase sous unité 1)

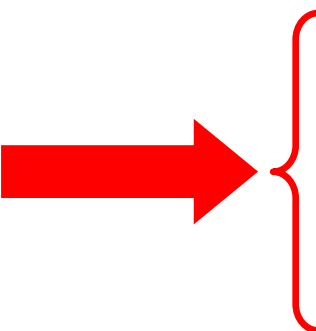


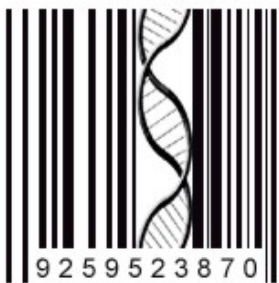


1. Le projet "code barre ADN" Pourquoi un gène mitochondrial?



- Beaucoup de copies (100-10000 copies d'un gène mt vs 2 copies pour le nucléaire)
- Absence d'introns
- Plus de différences **inter**-spécifiques que les gènes nucléaires (5x-10x).
- Moins de variabilité **intra**-spécifique

- 
- Séquencage plus facile
 - Plus d'info dans une courte séquence
 - Espèces mieux discriminées



1. Le projet "code barre ADN" Uniquement COI?

COI semble marcher pour beaucoup d'espèces animales

Pas assez variable pour les plantes, plusieurs autres gènes sont en cours d'étude :

- ITS, plusieurs séquences mitochondriales, *rbcL* (chloroplaste)

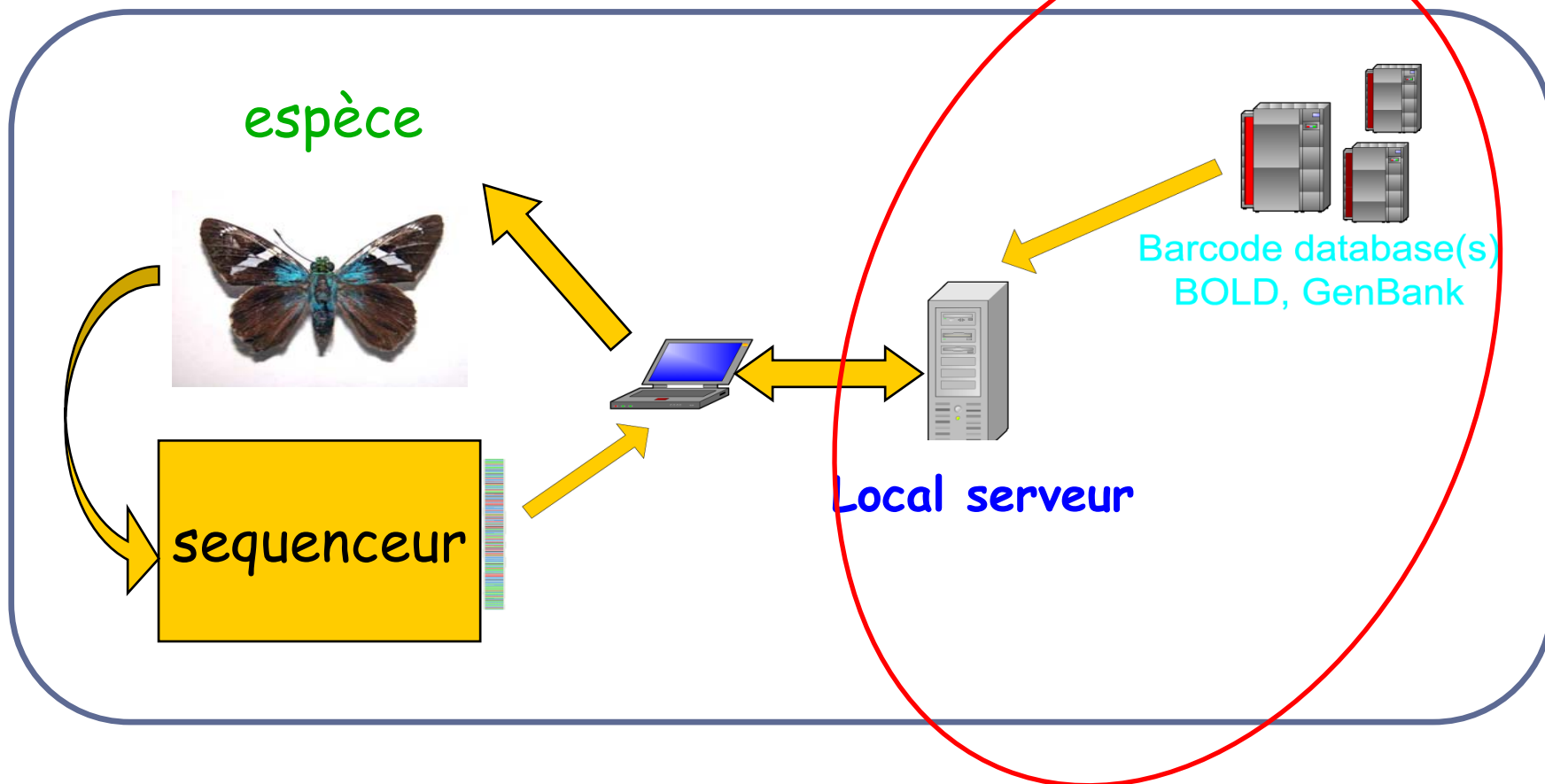
De même COI marche mal pour les amphibiens (trop forte variation)

- 16S

Et sans doute d'autres...



1. Le projet "code barre ADN" Qu'est ce que c'est?



Base de donnée globale regroupant toutes
les données associées "code barre ADN"

SPECIMEN DATA -Profile - Guelph Moths

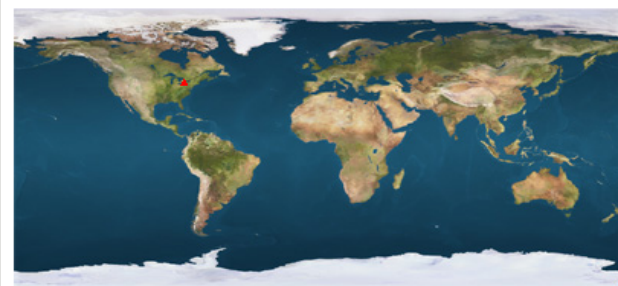
Identification : *Acronicta hastulifera*

Specimen Accession : moth682.01
Specimen Label : moth682.01

Sex :
Reproduction :
Life Stage :

GPS Latitude : 43.537
GPS Longitude : -80.1353
Elevation (meters) : 320

Country : Canada
State/Province : Ontario
Region : Wellington County
Sector : Puslinch Township
Site : Concession 11



▲ Collection Location [click on image to zoom]

GenBank Accession : [AF549746](#)

Museum Accession :
Institution Holding : University of Guelph
Collector : Paul Hebert
Date Collected : Unknown
Identifier : Paul Hebert

Common Name :
Taxonomy :
phylum - Arthropoda
class - Insecta
order - Lepidoptera
family - Noctuidae
subfamily - Acronictinae
genus - *Acronicta*
species - *Acronicta hastulifera*

Notes :



Wing Span = 5.5 cm

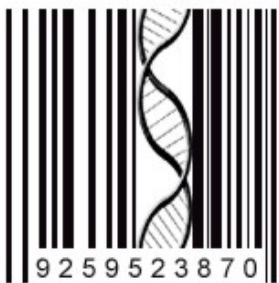
[Click on image to view images at larger size](#)

Liens avec les séquences dans GenBank section BarCoding



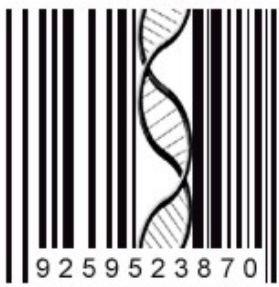
1. Le projet "code barre ADN" Les bénéfices (1)

- OK avec peu de tissus (frangments, poils, tissus mal conservés)
- OK pour tout stade de vie (ex: stades larvaires)
- Séquencage rapide et peu couteux (qq €, qq heures, beaucoup moins dans le futur proche cf "GPS barcoder")
- Identification de nombreux spécimen en un temps court
- Une fois la base de données établie, utilisable par des non-spécialistes



1. Le projet "code barre ADN" Les bénéfices/applications (2)

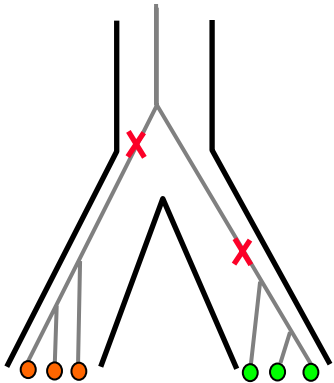
- Outils d'identification taxonomique, facile d'utilisation, complémentaire de la morphologie (biomedecine, agronomie, douanes,...)
- Facilite/accélère inventaires/suivis de biodiversité
- "Démocratise" l'accès aux résultats des études de systématique
- Peut faciliter la visibilité des collections
- Peut libérer les taxonomistes des tâches d'identifications simples -> plus de temps pour délimitation des espèces et découvertes de nouvelles espèces



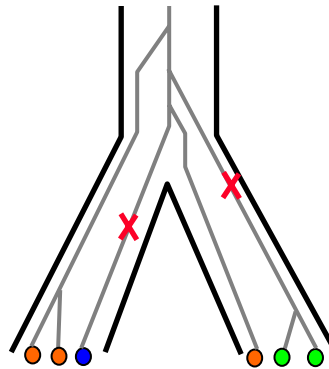
1. Le projet "code barre ADN"

Les problèmes et limites de l'approche (1)

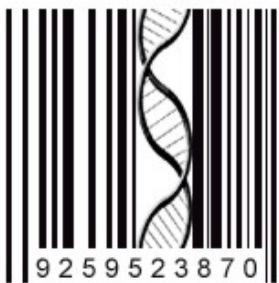
- Il n'existe pas de gène "universel/ideal" qui serait toujours variable entre espèce et invariant au sein de toutes espèces



Cas du gène idéal



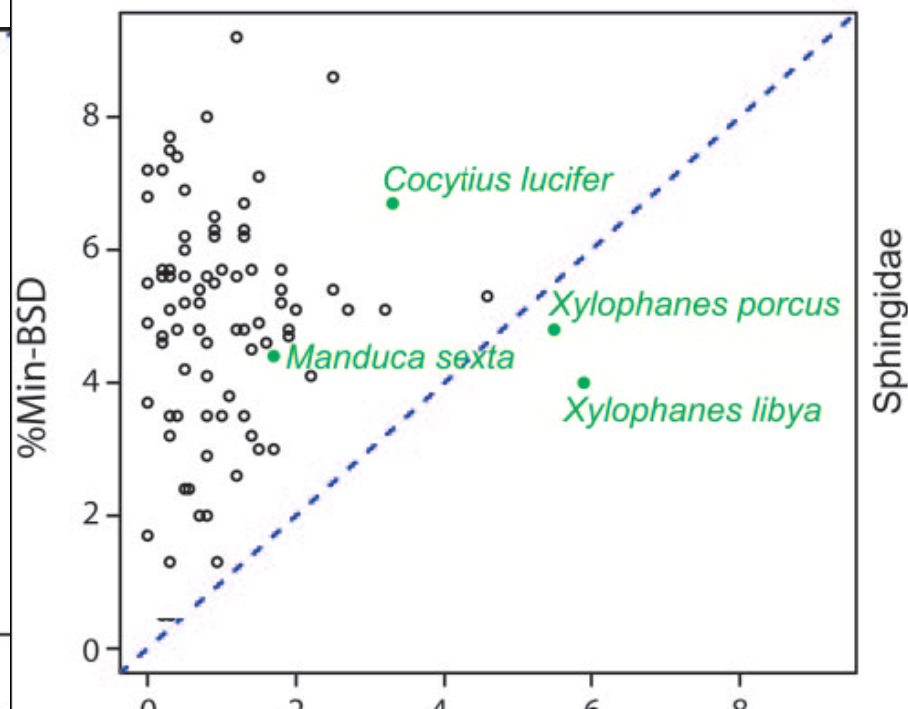
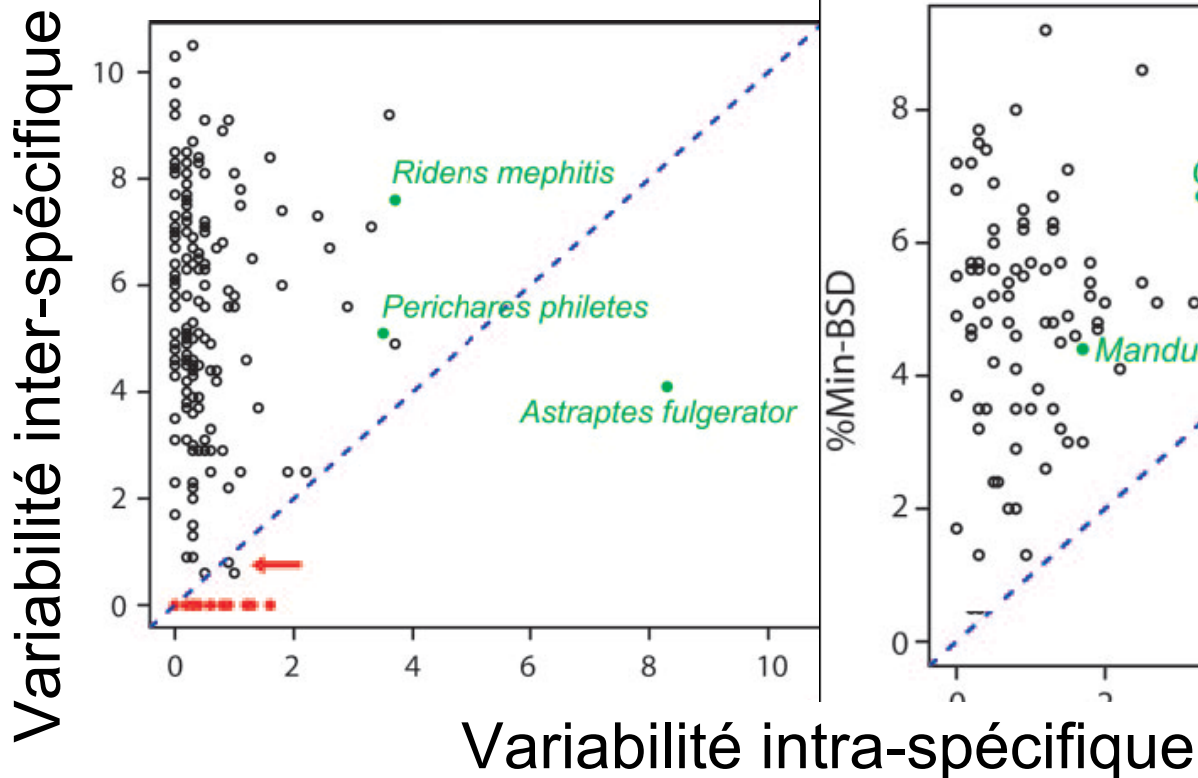
Mais la réalité est plus complexe

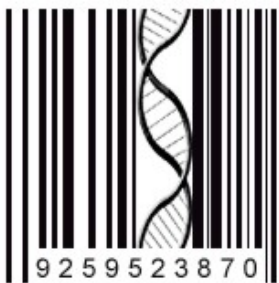


1. Le projet "code barre ADN"

Les problèmes et limites de l'approche (2)

- Plus globalement, il n'est pas toujours vrai que la variabilité intra-spécifique est faible par rapport à la variabilité inter-spécifique, même avec COI



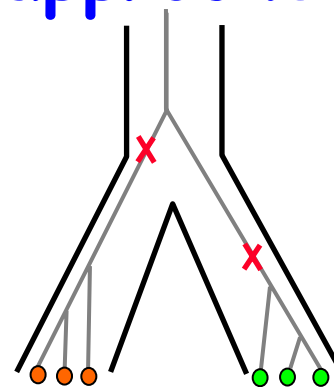


1. Le projet "code barre ADN"

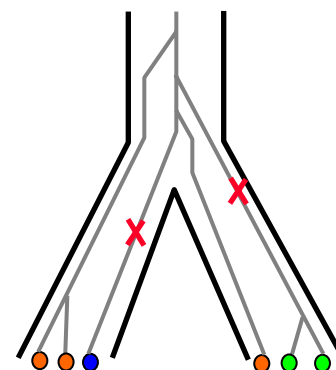
Les problèmes et limites de l'approche (3)

- Il n'existe pas de gène "universel/ideal" qui serait toujours variable entre espèce et invariant au sein de toutes espèces
- Plus globalement, il n'est pas toujours vrai que la variabilité intra-spécifique est faible par rapport à la variabilité inter-spécifique, même avec COI

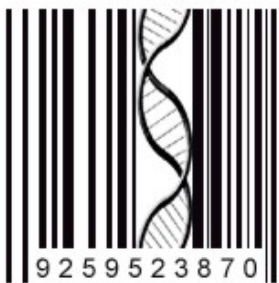
-> nécessité de développer un cadre d'analyse rigoureux permettant de séparer variabilité intra et inter-spécifique et de préciser l'incertitude de chaque détermination



Cas du gène idéal



Mais la réalité est plus complexe



1. Le projet "code barre ADN"

Les problèmes et limites de l'approche (4)

Ce n'est pas une nouvelle méthode pour faire de la taxonomie et de la systématique mais juste un outils de détermination :

- Les séquences de la base de données doivent être générée à partir de spécimens préalablement déterminés (muséums)

-> L'identification par "code barre AND" ne se fera que sur des groupes pour lesquels la taxonomie est déjà connue

Pourra seulement suggérer de ré-analyser selon les approches classiques de taxonomie certains taxons pour lesquels le "code barre ADN" suggérerait une séparation d'une espèce en plusieurs ou le regroupement de plusieurs espèces en une seule...



1. Le projet "code barre ADN"

Les problèmes et limites de l'approche (5)

Dans certains cas, la détermination par "code barre ADN" risque d'être difficile :

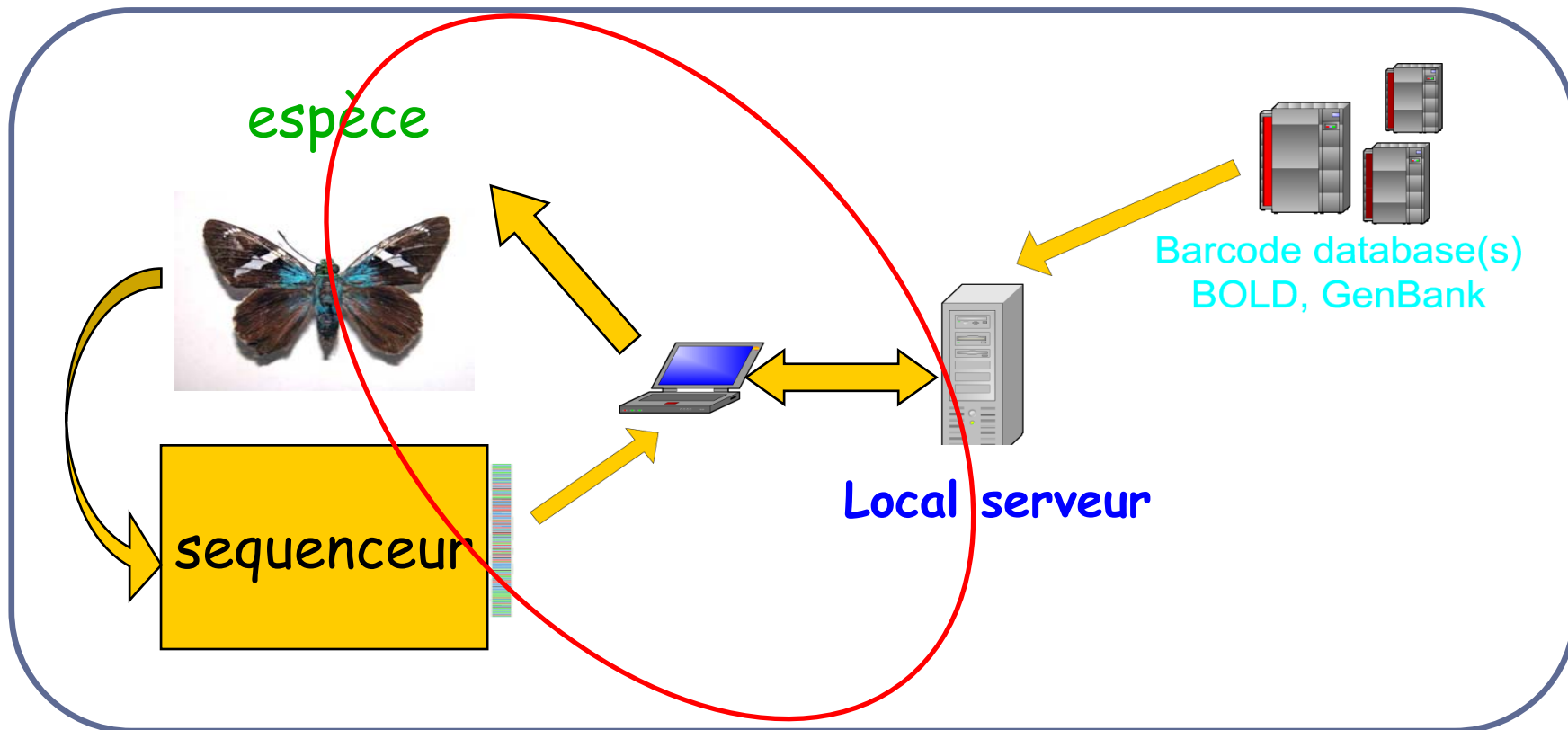
- Groupes montrant une très faible ou très forte diversité
- Espèces fortement structurées dans l'espace
- Espèces ayant divergées récemment
- Toutes populations/espèces montrant de l'hybridation/introgression

Définir des marqueurs supplémentaires et un échantillonnage plus poussé pour résoudre ces problèmes :

- Marqueurs plus variables (divergence récentes, hybridation, faible diversité) : séq. hypervariables, microsatellites
- Marqueurs nucléaires recombinants (hybridation, forte diversité)
recombinaison = même espèce
- Echantillonnage sur plusieurs aires géographiques



1. Le projet "code barre ADN" méthodes d'analyse des données



Outils d'analyse :

Permet d'interroger la base de données et de proposer une espèce à partir de la séquence présentée



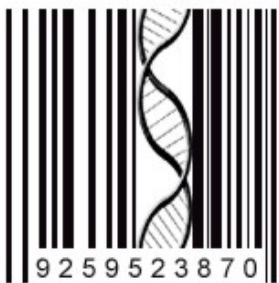
1. Le projet "code barre ADN" quelles méthodes d'analyse?

But = assignation d'une séquence à un groupe taxonomique
génétiquement caractérisé

= distinguer variabilité intra- vs. inter-spécifique

4 grandes classes de méthodes :

1. Similitudes/Appariement de séquences
2. Classification statistiques
3. Phylogénétiques (Neighbors-joining, maximum vraisemblance)
4. Modèle populationnels : Spéciation + Coalescence
(maximum vraisemblance, Bayésien)



1. Le projet "code barre ADN" quelles méthodes d'analyse?

1. Similitudes/Appariement de séquences : Blast, Google gene

L'espèce est donnée par la séquence de la base de données la plus proche de la séquence focale

+ rapidité

- pas de distinction variabilité intra-/inter-spé
- pas de signification/modèle biologique
- pas de résolution des ambiguïtés

Ca marche quand même dans >70% des cas testés!

NNAACATTATATTTTATTTTGGAAATTTGAGCAGGAATAGTTGGAACCT
CACTAAGATTACTAATTGAGCAGAA

GoogleGene Search

Clear

21 bases = 1 character (word)

>Query sequence [651 bases (31 characters) out of 660 original bases]

NNAACATTATATTTTATTTT GGAATTTGAGCAGGAATAGTT GGAACCTCACTAAGATTACTA
ATTGAGCAGAAATTAGGAACC CCCGATCTTTAATTGGAGAT GACCAAAATTTATAACACAATT
GTTACAGCTCATGCATTTATT ATAATTTTTTTTATAGTAATA CCAATTATAATTGGAGGATTT
GGTAATTGATTAGTACCTTTA ATATTAGGAGCACCTGATATA GCATTCCCACGAATAAATAAC
ATAAGATTTTGACTTTTACCC CTTTCATTAACTCTTTTAATT TCTAGAAGTATTGTAGAAAAC
GGAGCAGGAACCTGGTTGAACA GTTTACCCCTCTCTCTCTCTT AACATTGCTCATAGTGGAACT
TCTGTAGATTTAGCTATTTT TCCCTTCATTTAGCTGGTATT TCTTCAATTATAGGAGCTGTA
AATTTTATTACTACTATTATT AATATGCGAATTAATAATTTA TCATTTGATCAAATACCATT
TTGTTTGAGCTGTTGGAATC ACAGCCTTTTTATTATTACTA TCTTTACCAGTATTAGCTGGT
GCAATTACAATATTATTAAC GATCGAAATCTTAATACATCA TTTTTTGACCTGCTGGAGGG
GGAGACCTTATTCTATATCAA

Sequences matching your query:

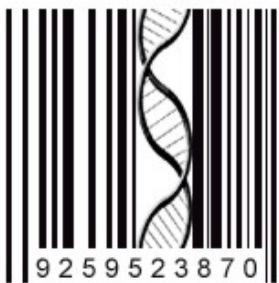
>MHASC043 05 02 SRNP 19434 Xylophanes Janzen01 [31 out of 31 characters match]

NNAACATTATATTTTATTTT GGAATTTGAGCAGGAATAGTT GGAACCTCACTAAGATTACTA
ATTGAGCAGAAATTAGGAACC CCCGATCTTTAATTGGAGAT GACCAAAATTTATAACACAATT
GTTACAGCTCATGCATTTATT ATAATTTTTTTTATAGTAATA CCAATTATAATTGGAGGATTT
GGTAATTGATTAGTACCTTTA ATATTAGGAGCACCTGATATA GCATTCCCACGAATAAATAAC
ATAAGATTTTGACTTTTACCC CTTTCATTAACTCTTTTAATT TCTAGAAGTATTGTAGAAAAC
GGAGCAGGAACCTGGTTGAACA GTTTACCCCTCTCTCTCTT AACATTGCTCATAGTGGAACT
TCTGTAGATTTAGCTATTTT TCCCTTCATTTAGCTGGTATT TCTTCAATTATAGGAGCTGTA
AATTTTATTACTACTATTATT AATATGCGAATTAATAATTTA TCATTTGATCAAATACCATT
TTGTTTGAGCTGTTGGAATC ACAGCCTTTTTATTATTACTA TCTTTACCAGTATTAGCTGGT
GCAATTACAATATTATTAAC GATCGAAATCTTAATACATCA TTTTTTGACCTGCTGGAGGG
GGAGACCTTATTCTATATCAA CATTATTTT

>MHASC038 05 02 SRNP 18238 Xylophanes Janzen01 [30 out of 31 characters match]

NNAACATTATATTTTATTTT GGAATTTGAGCAGGAATAGTT GGAACCTCACTAAGATTACTA
ATTGAGCAGAAATTAGGAACC CCCGATCTTTAATTGGAGAT GACCAAAATTTATAACACAATT
GTTACAGCTCATGCATTTATT ATAATTTTTTTTATAGTAATA CCAATTATAATTGGAGGATTT
GGTAATTGATTAGTACCTTTA ATATTAGGAGCACCTGATATA GCATTCCCACGAATAAATAAC
ATAAGATTTTGACTTTTACCC CTTTCATTAACTCTTTTAATT TCTAGAAGTATTGTAGAAAAC
GGAGCAGGAACCTGGTTGAACA GTTTACCCCTCTCTCTCTT AACATTGCTCATAGTGGAACT
TCTGTAGATTTAGCTATTTT TCCCTTCATTTAGCTGGTATT TCTTCAATTATAGGAGCTGTA
AATTTTATTACTACTATTATT AATATGCGAATTAATAATTTA TCATTTGATCAAATACCATT
TTGTTTGAGCTGTTGGAATC ACAGCCTTTTTATTATTACTA TCTTTACCAGTATTAGCTGGT
GCAATTACAATATTATTAAC GATCGAAATCTTAATACATCA TTTTTTGACCTGCTGGAGGA
GGAGACCTTATTCTATATCAA CATTATTTT





1. Le projet "code barre ADN" quelles méthodes d'analyse?

2. Modèles statistiques de classification

Etape1 : Recherche de critères de classification (= les nucléotides les plus discriminants ou des distances génétiques) pour former des groupe homogènes (= les espèces) à partir de sur toutes les séquences de la base de données

Etape2 : Classer la séquence focale selon les mêmes critères dans un des groupes

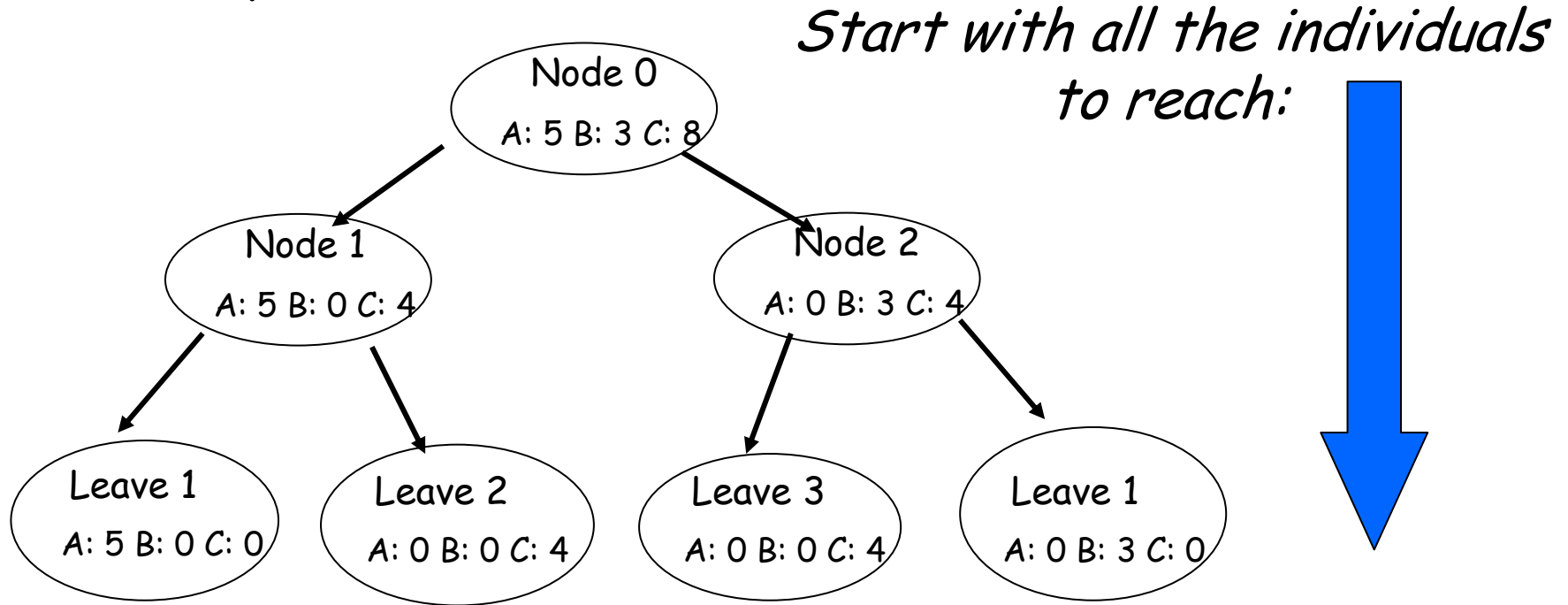
- + meilleure classification
- + incertitude possible
- parfois lent
- pas de signification/modèle biologique

Marche dans >80% des cas

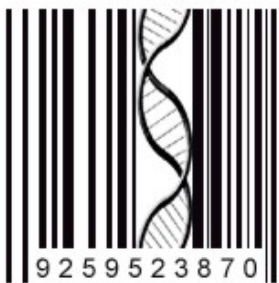
Ex: CART (Classification And Regression Tree)

(Breiman et al., 1984, 1996)

- Builds a classification tree from the reference sample (= data base)



Then assign the unknown sequence by applying the same rules

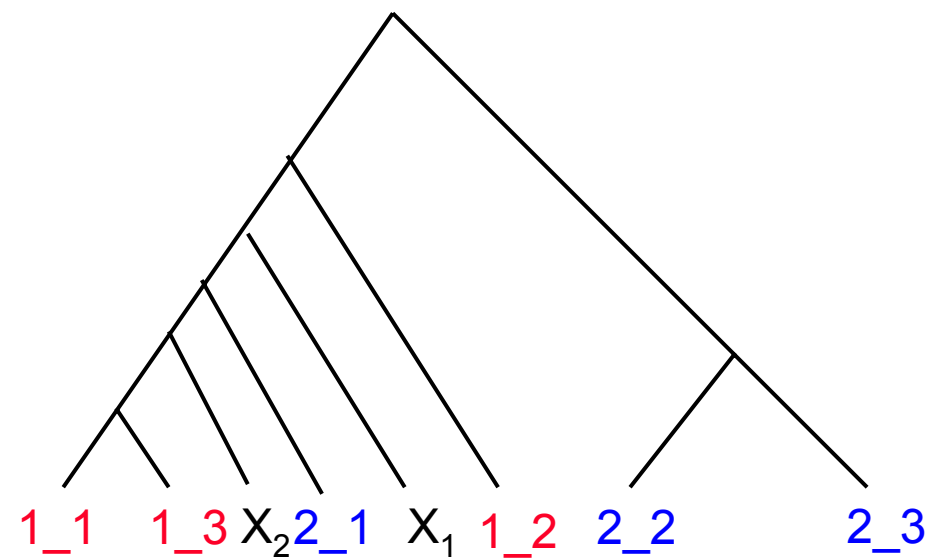


1. Le projet "code barre ADN" quelles méthodes d'analyse?

3. Phylogénétiques (neighbor joining, maximum vraisemblance)

Construction d'un arbre phylogénétique à partir des séquences de la base de données + la séquence focale

Especie = là ou se place la séquence focale par rapport aux autres, mais pas vraiment de critère universel



Règles de décision possibles :

- Unanimité : X1 classé comme ambigu, X2 classé comme espèce 1
- Majorité : X1 and X2 classés comme espèce 1



1. Le projet "code barre ADN" quelles méthodes d'analyse?

3. Phylogénétiques (neighbor joining, maximum vraisemblance)

- + relativement rapide
- + prendre en compte des modèles d'évolution moléculaire
- + cadre d'analyse de données génétique connu
- + marche mieux que les autres quand polymorphisme fort
- Marche moins bien que les autres quand polymorphisme faible
- Pas de critère d'incertitude (juste ambiguïtés)

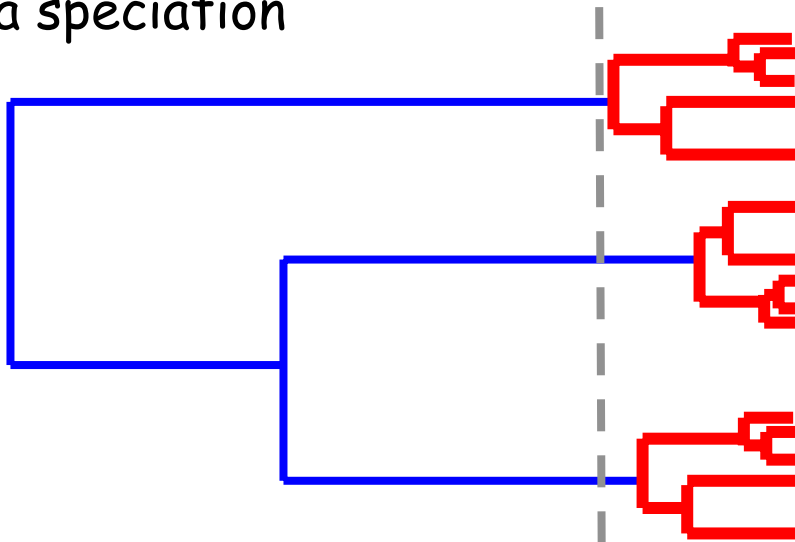
Marche mieux que les méthodes précédentes
quand forte variabilité intra-spécifique



1. Le projet "code barre ADN" quelles méthodes d'analyse?

4. Modèle populationnels : Spéciation + Coalescence

Extension des approches de génétique des populations pour prendre en compte la spéciation



Branchements inter-espèce
Modèles de spéciation :
Taux de spéciation, d'extinction

Branchements intra-espèce
Théorie de la coalescence :
Taille de pops, flux de gènes,
Histoire démographique et selective

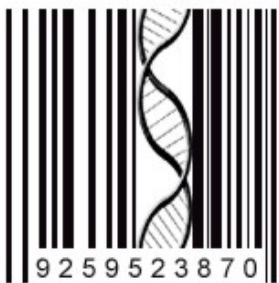


1. Le projet "code barre ADN" quelles méthodes d'analyse?

4. Modèle Spéciation + Coalescence = extension génétique des pops

Peu d'applications "code barre ADN" développées jusque là mais :

- + Cadre statistique optimal (max vraisemblance)
 - > utilisation de toute l'info des données, incertitude, test hyp....
- + Cadre theorique très développé, bien connu et **puissant** (coalescent)
- + Signification biologique forte, modèles populationnels réalistes
- + Inférence biologiques possibles (limitées ici car 1 seul ou peu de gènes)
- Fondées sur marqueurs neutres (mais certaines selections +- OK)
- Développée pour multilocus, ici un seul ou peu de gènes
 - > moindre puissance
- Demande des moyen de calcul importants si modèles complexes



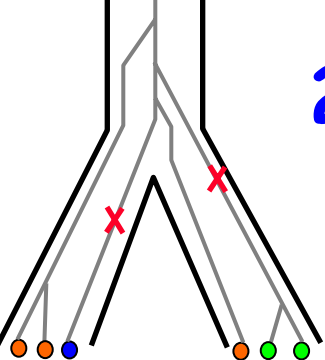
1. Le projet "code barre ADN" quelles méthodes d'analyse?

4. Modèle Spéciation + Coalescence = extension génétique des pops

Peu d'applications "code barre ADN" développées jusque là mais :

- Devrait marcher quand les autres méthodes ne marche pas
- Nécessite une bonne description de la variabilité intra-spécifique
(-> échantillonnage intra-spé important)
- Assignment lente mais la plus précise : Utilisation pour l'étape finale
d'assignment à une espèce quand elle est problématique

A tester ...

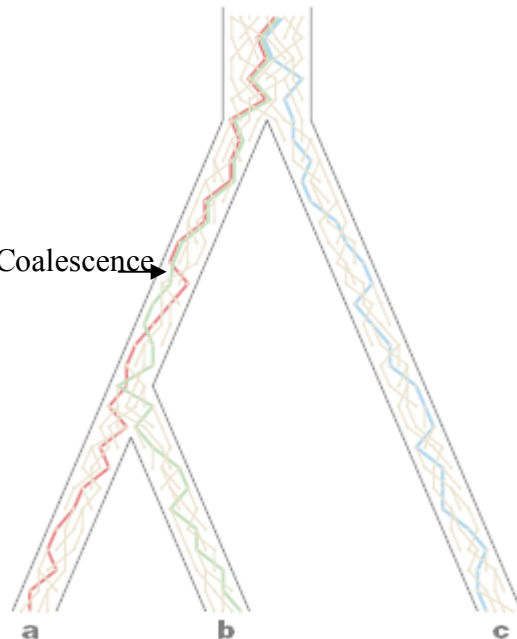
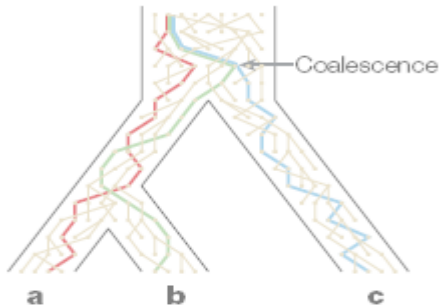


2. Arbres d'espèces et arbres de gènes ="phylogénie et coalescence"

Histoire généalogique d'un gène = processus stochastique de forte variance (coalescent)

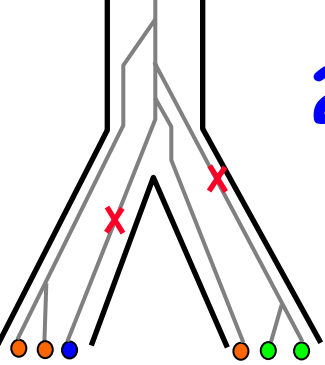
-> dans quelle mesure l'arbre d'un gène et l'arbre des espèces sont concordants???

= tri des lignées ancestrales



Concordants si :

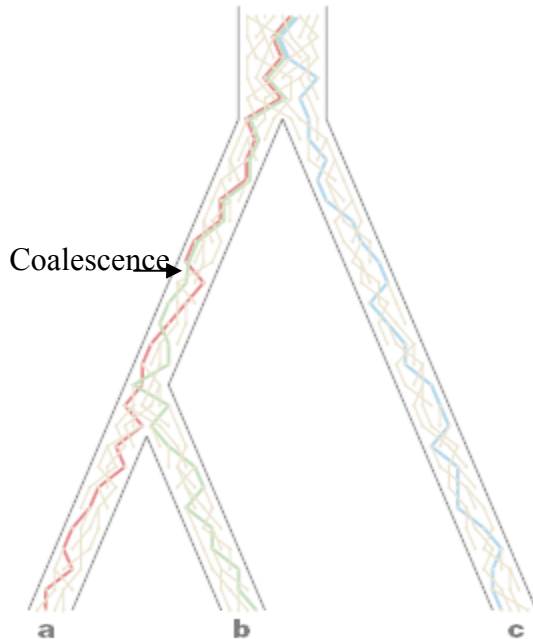
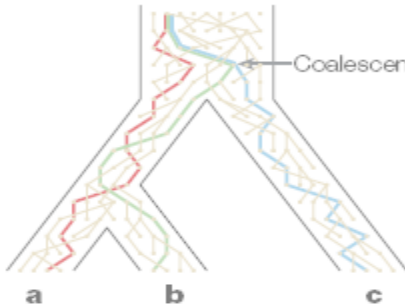
- Temps spéciation \gg temps de coalescence
- Pas/peu de flux de gènes pendant l'événement de spéciation
- Pas de transfert de gènes horizontaux



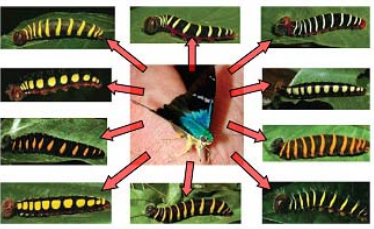
2. Arbres d'espèces et arbres de gènes ="phylogénie et coalescence"

Concordants si :

• Temps spéciation \gg temps de coalescence



Fortement influencé
par des facteurs populationnels,
Temps de coalescence courts si :
Petites tailles de populations
Sélection



3. Besoin d'une évaluation du projet "code barre ADN"

(1) les cas problématiques :

- Divergence récente
- Hybridation/Introgression
- Diversité très faible ou très forte
- Grandes populations
- Populations fortement structurées dans l'espace

Notamment isolement par la distance (dispersion localisée)

+ cas spécifiques : ex. dispersion femelles \neq dispersion mâles

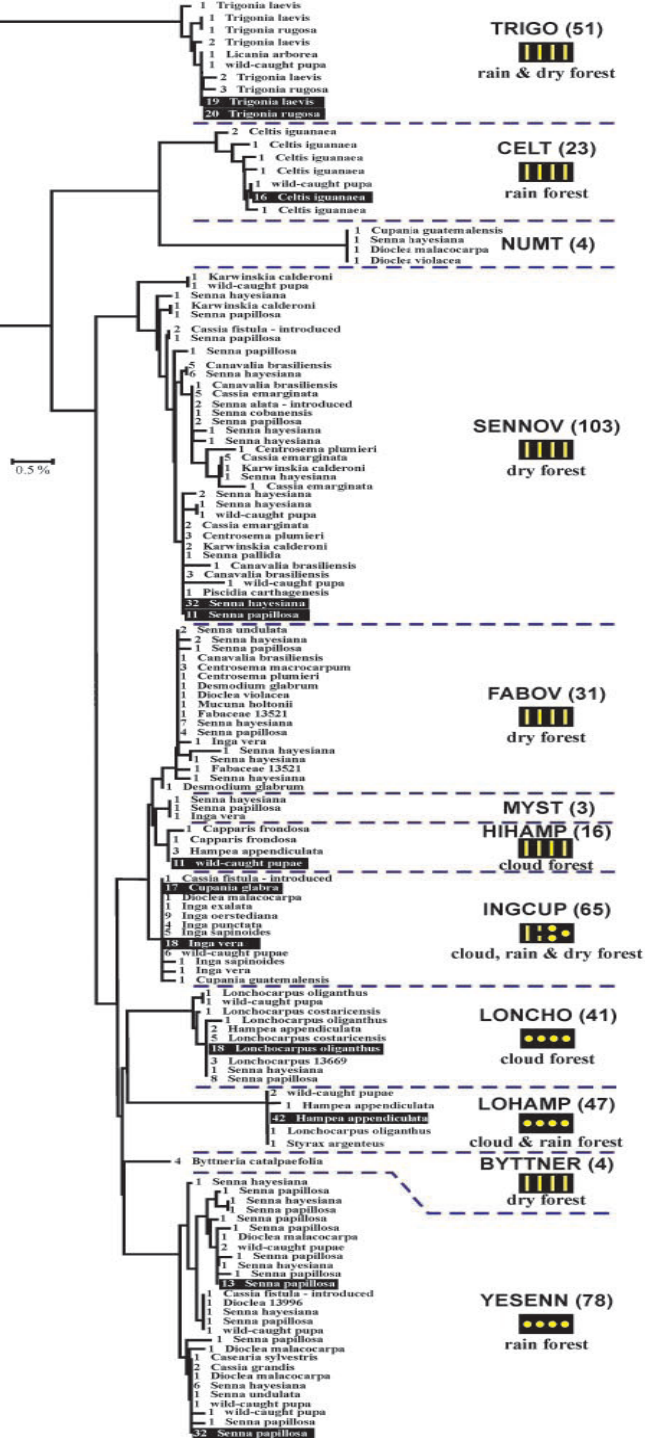
(cf exemple *Astraptes* développé ci après)

Un exemple de cas problématique :

**Ten species in one: DNA barcoding reveals
cryptic species in the neotropical skipper
butterfly *Astraptes fulgerator***

Paul D. N. Hebert*†, Erin H. Penton*, John M. Burns‡, Daniel H. Janzen§, and
Winnie Hallwachs§

PNAS October 12, 2004 vol. 101 no. 41 14812–14817



TRIGO



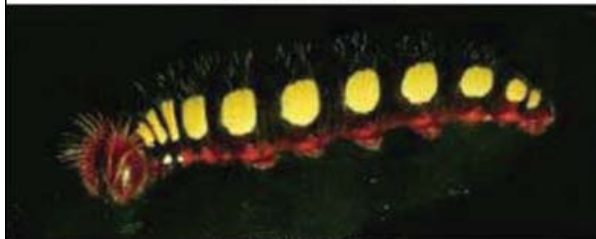
CELT



LONCHO



INGCUP



LOHAMP



HIHAMP



BYTTNER



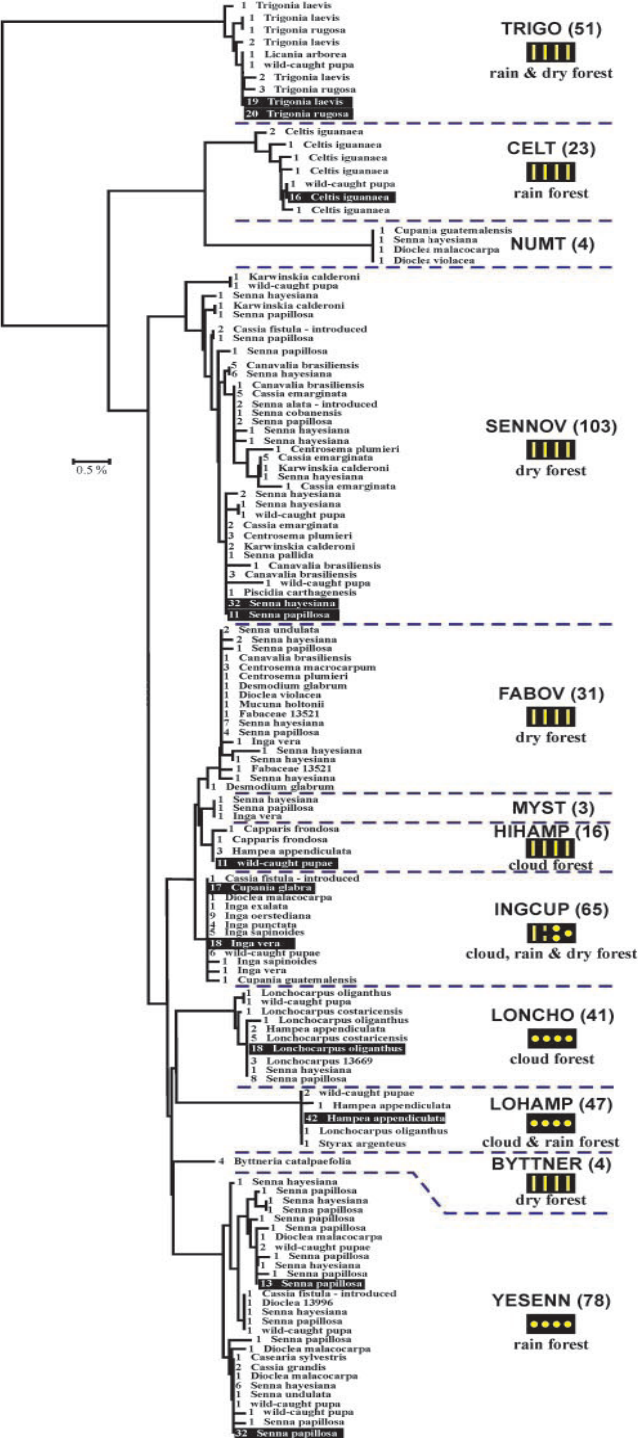
FABOV



YESENN



SENNOV



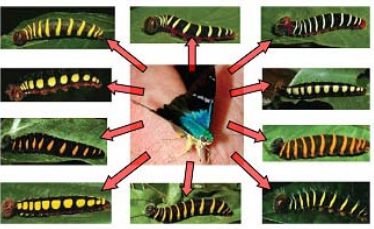
→ dans 1 espèce (déjà décrite et étudiée), ils en trouvent 10!!

Chacune de ces 10 espèces serait associée à une famille spécifique de plantes

Problème potentiel : Femelles choisissent leur lieux de ponte (= une plante) et ADN mt transmit exclusivement par femelles

-> si les femelles ne change pas/peu de milieu de ponte mais que les mâles se reproduisent avec toutes les femelles sans préférences, on aura un signal de structuration sur l'ADN mt qui ne se retrouvera pas sur l'ADN nucléaire...

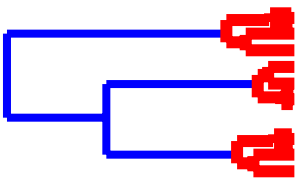
Intéressant mais pas forcément 10 espèces différentes...



3. Besoin d'une évaluation du projet "code barre ADN"

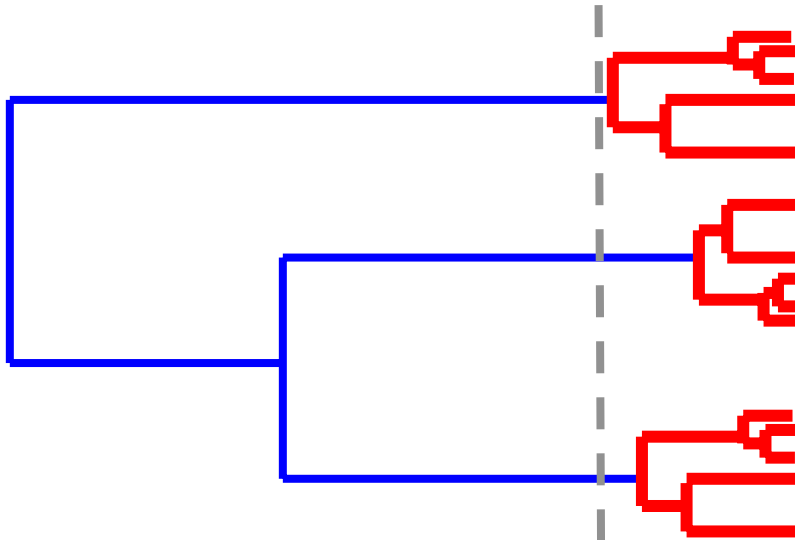
(2) Comment améliorer les choses:

- Echantillonnage intra-spécifique plus poussé (>3 ind/esp. = 10/15)
- Séquencage d'autres gènes
 - gènes nucléaires recombinants
 - gènes plus variables (microsats pour spéciation récente?)
- Prendre en compte la structuration géographique :
 - Comment adapter l'échantillonnage?
- Effet des fluctuations démographiques passées
 - Méthodes robustes sinon les prendre en compte (problématique...)
- Comment gérer les espèces non échantillonnées ou inconnues?



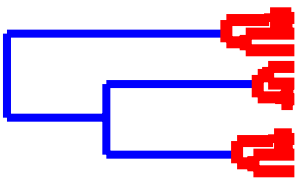
4. Besoin de nouvelles méthodes d'analyse

Extension des approches populationnelles à l'étude des espèces : Coalescence + spéciation



Branchements inter-espèce
Modèles de spéciation :
Taux de spéciation, d'extinction

Branchements intra-espèce
Modèles populationnels
Théorie de la coalescence :
Taille de pops, flux de gènes,
Histoire démographique et selective



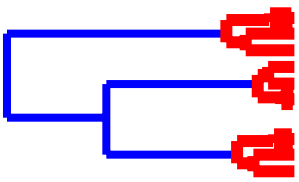
4. Besoin de nouvelles méthodes d'analyse

Extension des approches populationnelles à l'étude des espèces : Coalescence + spéciation

Quelques résultats encourageants :

Coalescent assigner					distance method				
$n = 5$	$\nu \setminus \theta$	0.001	0.01	0.1	$n = 5$	$\nu \setminus \theta$	0.001	0.01	0.1
	0.001	93%	89%	82%		0.001	92%	61%	56%
	0.01	99%	97%	92%		0.01	99%	93%	60%
	0.1	100%	100%	99%		0.1	100%	100%	92%
	1	100%	100%	100%		1	100%	100%	99%

Abdo & Golding, 2005, Data analysis working group meeting,

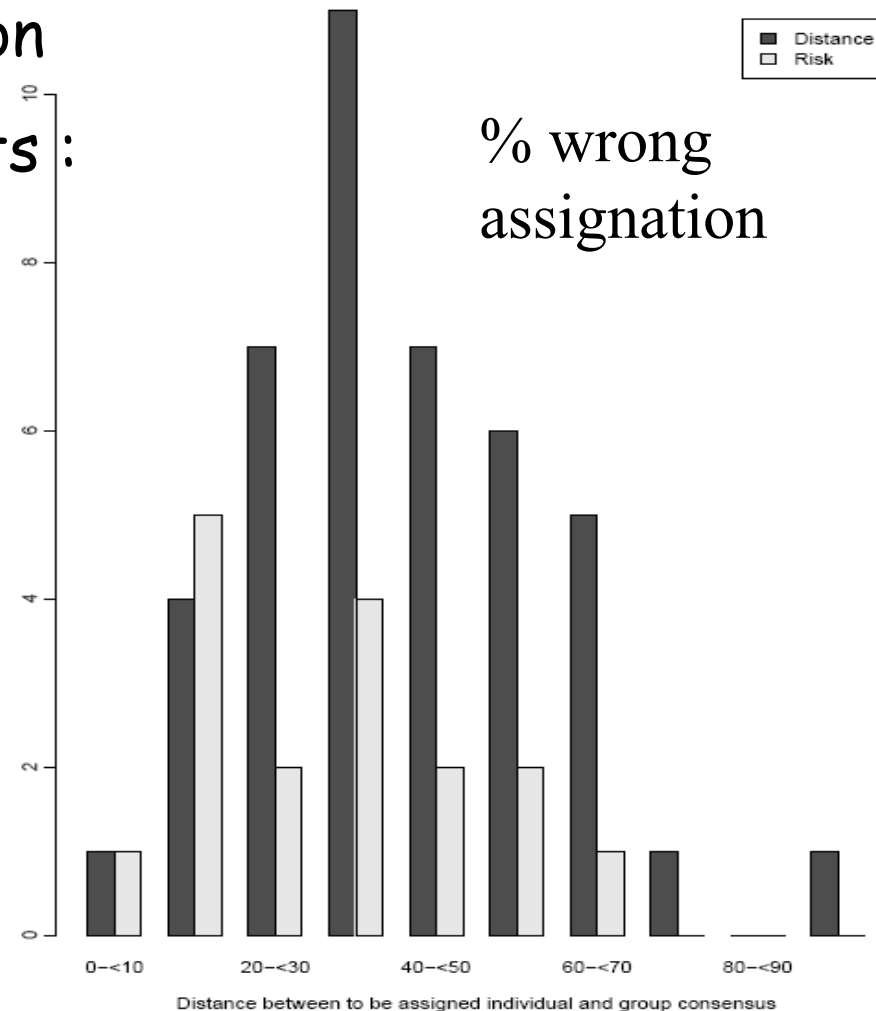


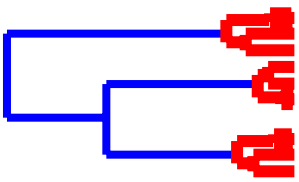
4. Besoin de nouvelles méthodes d'analyse

Extension des approches populationnelles à l'étude des espèces : Coalescence + spéciation

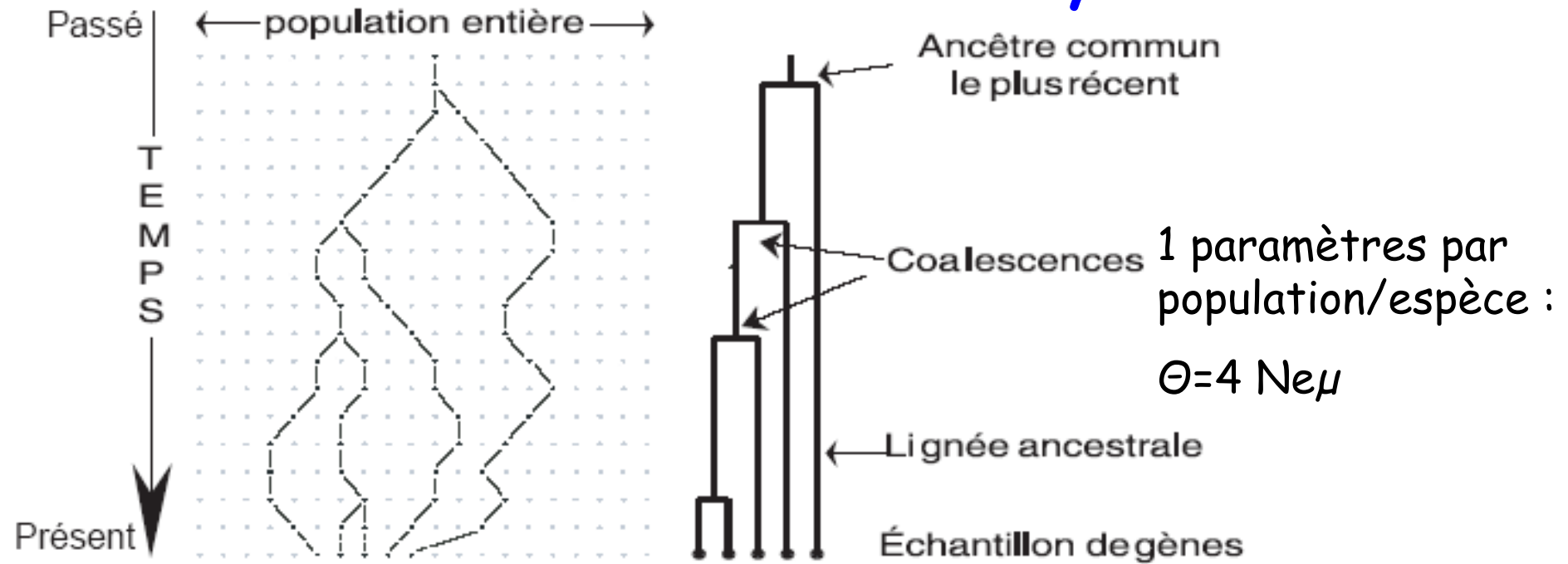
Quelques résultats encourageants :

Abdo & Golding, 2005, Data analysis working group meeting,



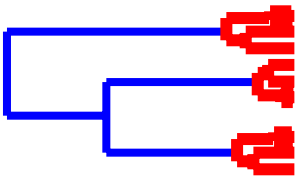


4. Besoin de nouvelles méthodes d'analyse



Autres intérêts de la coalescence :

- Structuration géographique : facilement incorporable (mise à l'échelle $\Theta = 4 \alpha N_e \mu$)
- Fluctuations démographiques : peut être pris en compte
- Espèces/pops non échantillonnées ou inconnues : facilement incorporable

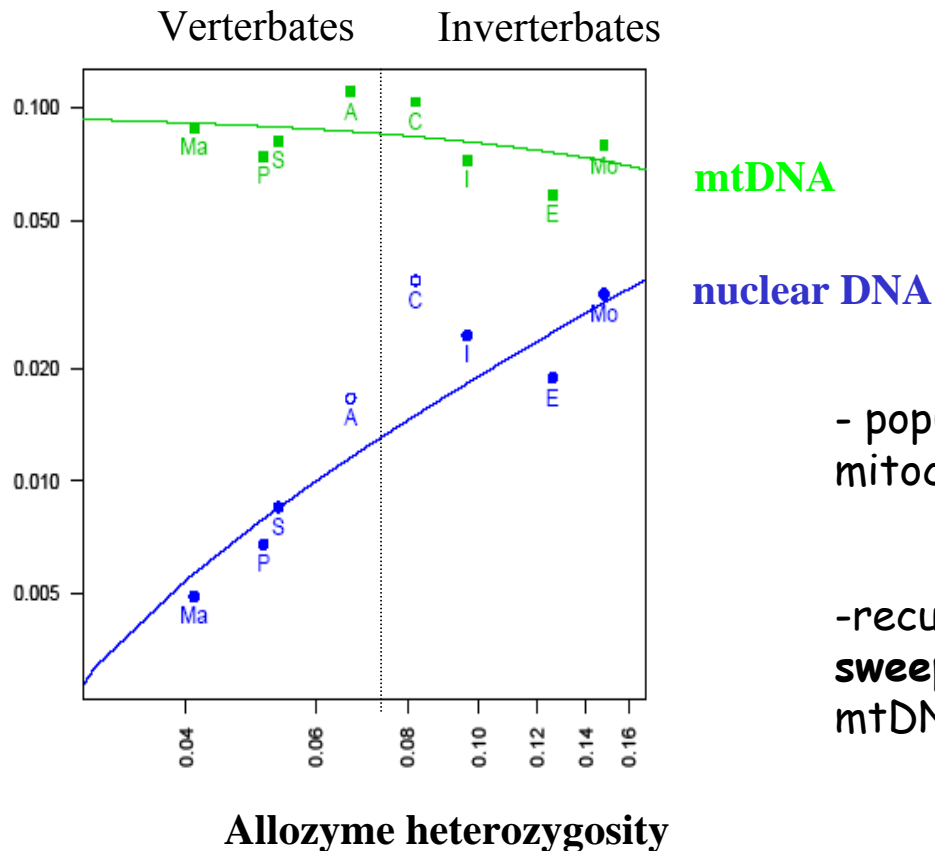


4. Besoin de nouvelles méthodes d'analyse

Les problèmes des modèles Coa + spéciation :

- Tx de spéciation variables + Θ avec mise à l'échelle variables,
 - > beaucoup de paramètres de nuisance mais possibilité de les estimer uniquement avec la base de données de référence, réutilisable ensuite pour l'assignation de séquences focales
- Coalescent = modèle neutre, COI fonctionne-t-il comme un gène neutre??

Bazin et al. 2006 Science

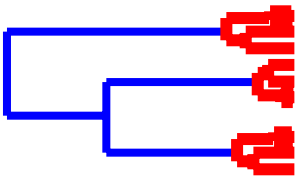


- population size influences nuclear, but not mitochondrial DNA diversity

-recurrent adaptive evolution (**selective sweeps**) can explain the homogeneous mtDNA pattern

Peut on utiliser la coalescence pour modéliser des gènes mitochondriaux?

Sans doute mais il faut tester si c'est robuste...



3. Besoin de nouvelles méthodes d'analyse

Les problèmes des modèles Coa + spéciation :

- Peu de locus mais on ne cherche pas estimer beaucoup de paramètres, uniquement tester l'appartenance d'une séquence à une espèce.
- temps de calcul long mais nouveau algorithmes plus efficaces pour modèles relativement simples



Conclusion

Le projet "code barre ADN" potentiellement intéressant mais beaucoup de points restent à tester pour que ça marche à 99%

->important de tester sur des jeux de données "ambigus"

Extension des approches populationnelles aux espèces

bon cadre pour quantifier, tester et prédire les erreurs et incertitudes de l'approche

perspective : utilisable pour étudier de façon combinée la micro et macro évolution et les mécanismes de spéciation, et la répartition intra- et inter- spécifique de la diversité génétique

Avec un peu d'imagination....



- A taxonomic GPS
- Link to reference database
- Usable by nonspecialists.

