Reconstruction de séquences ancestrales ou résurrection

Stéphanie Bertrand

stephanie.bertrand@obs-banyuls.fr

Qu'est-ce que c'est ?

•Séquence biologique (RNA, DNA, protéine..) d'un organisme aujourd'hui disparu

•Pour la majorité des cas, pas d'existence de séquence fossiles, on doit donc inférer les séquences ayant existé à partir de séquences existantes aujourd'hui

•Une des idées est un peu comme la paléonthologie: analyser ce qui est arrivé pour comprendre ce que l'on voit aujourd'hui

Pourquoi?

✓Phylogénie et comparaison fonctionnelles: inférences de la fonction ancestrale. Pas de modèle explicite d'évolution fonctionnelle

✓Reconstructions ancestrales: inférences sur la séquences et tests fonctionnels directs sur la protéine

✓Données sur les fonctions protéiques: données sur le milieu dans lequel les espèces ancestrales vivaient (biologie de la planète)

Pourquoi?

- On ne peut pas proposer de **modèle explicite** de l'évolution de la fonction précise d'une protéine: une mutation unique peut changer le substrat reconnu par une enzyme alors que 20 changements d'AA ne vont rien changer....
- Même en connaissant les sites importants pour l'activité d'une protéine (par exemple sites d'interaction récepteur/ligand déterminé par structure 3D), on ne peut déduire la fonction protéique uniquement à partir de sa séquence.
- La reconstruction de séquences ancestrales est un **outil** extrêmement **puissant** pour l'analyse de **l'évolution fonctionnelle**.

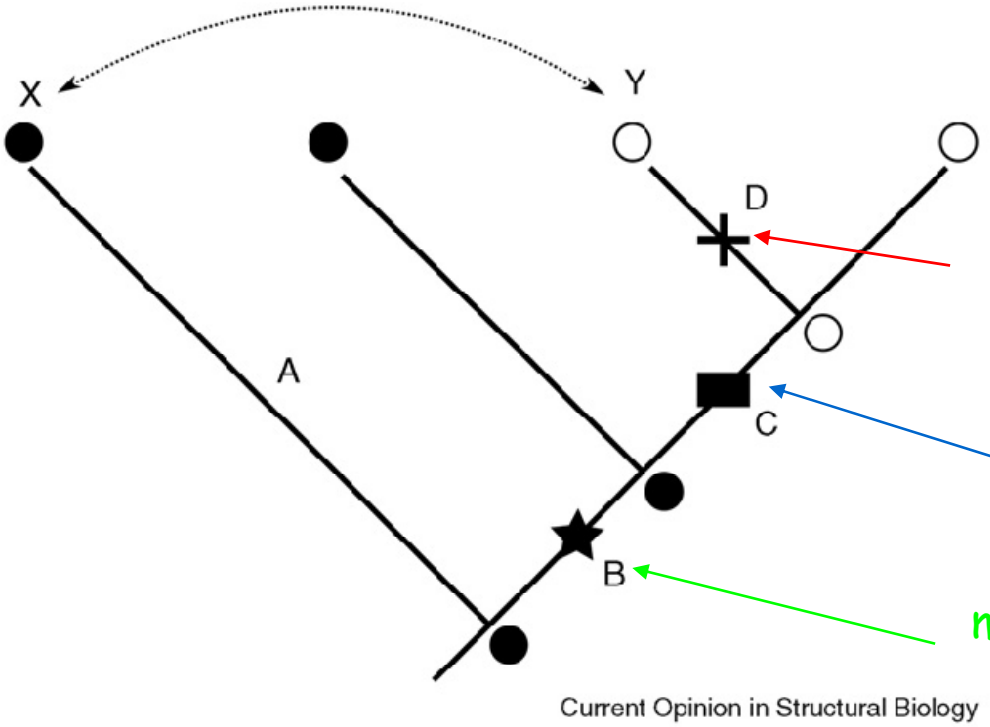
Même si ce n'est que de l'inférence et qu'il faut interpréter les résultats avec des « pincettes »……

Pourquoi?

Analyse de la relation séquence/structure/fonction des protéines

Analyse horizontale (comparaison X et Y) peut être non-concluante en raison de l'épistasie=interdépendance entre des mutations qui seules n'ont aucun effet ou des effets différents selon leur association à d'autres mutations

Analyse horizontale ne prend pas en compte l'histoire évolutive « cachée », pour la prendre en compte il faut remonter aux ancêtres.....

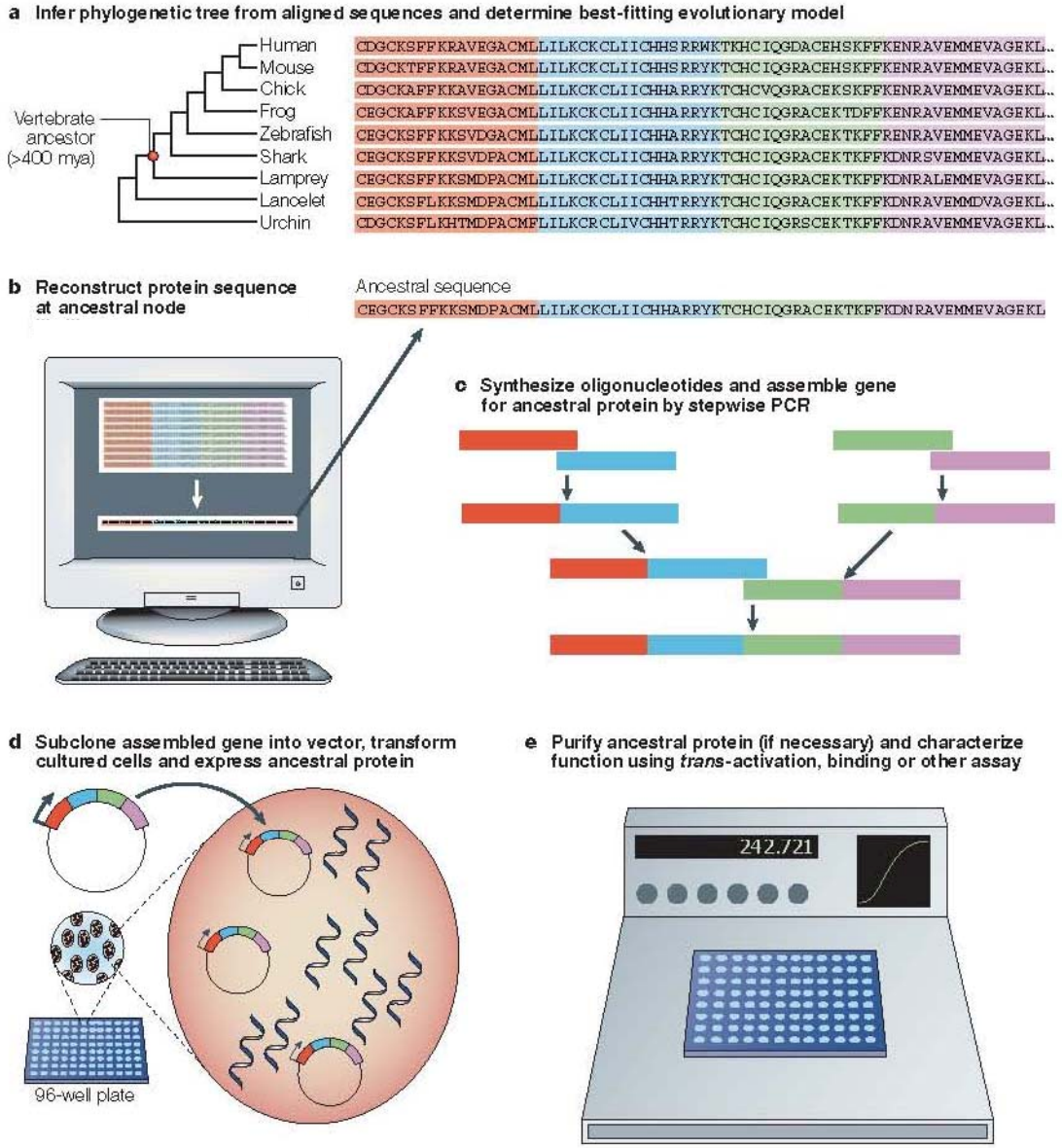


mutations incompatibles avec la fonction ancestrale

mutations permettant le changement de fonction

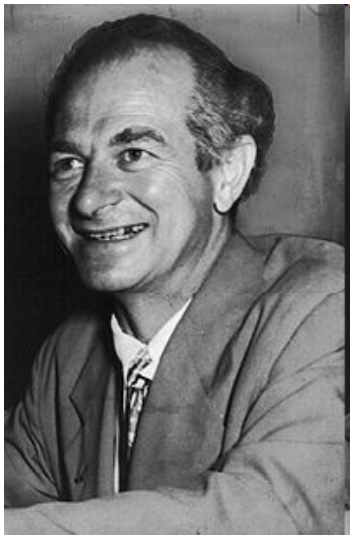
mutation permissive permettant à la protéine de « tolérer » le changement C

Comment faire?



From Thornton JW, Nature Reviews, 2004

Historique de la résurrection



Chemical Paleogenetics
Molecular « Restoration Studies » of Extinct Forms of Life
Acta Chemica Scandinavica 17 (1963) S9-S16

Les visionnaires: **1963**: **Linus Pauling et Emile Zuckerkandl** furent les premiers à proposer qu'un jour il serait possible d'inférer les séquences d'espèces éteintes pour « **synthétiser ces composants présumés d'organismes éteints...et étudier les propriétés physico-chimiques de ces molécules** »

Les pères du principe d'horloge moléculaire: « ...on the basis of this hypothesis, the degree of difference between two homologous polypeptide chains is the mesure of the relative time at which the common ancestor ... existed »

Historique de la résurrection

Le rêve ne devient réalité que grâce aux avancées

(i) des méthodes de phylogénie moléculaire

(ii) de l'ingénierie de l'ADN et de la production de protéines recombinantes

(iii) des possibilités de tests fonctionnels sur divers types de protéines

Premier article en **Mars 1990** in FEBS Letters:

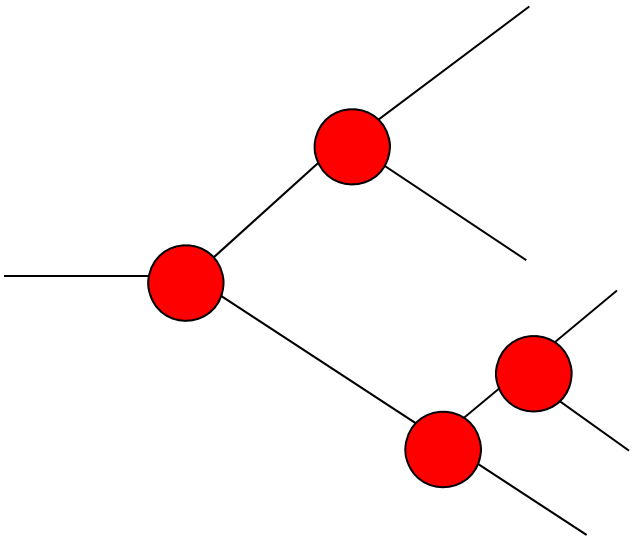
« The ribonuclease from an extinct bovid ruminant »

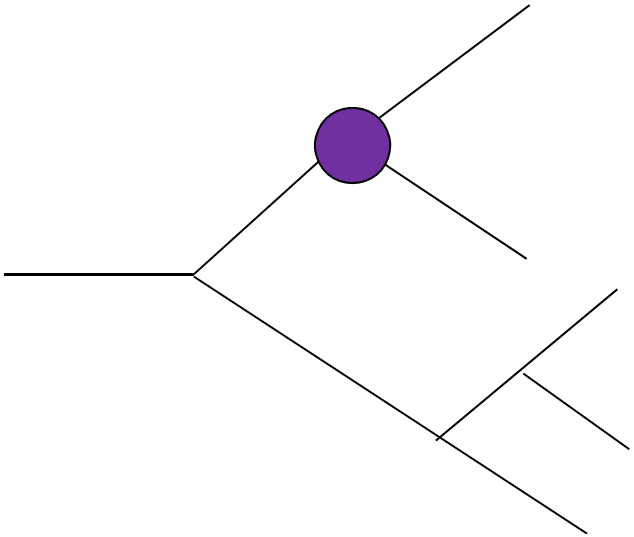
Stackhouse J et al.

Deux types de reconstruction:

Joint reconstruction (on s'intéresse aux ancêtres à plusieurs nœuds de l'arbre)
Marginal reconstruction (on s'intéresse à l'ancêtre à un nœud bien précis)

Dans les deux cas il faut:
•Un alignement des séquences actuelles
•Un arbre décrivant les relations entre ces séquences
•Un modèle d'évolution (pour méthodes probabilistes)





Différentes méthodes

Les différentes méthodes de reconstruction évoluent en même temps que les avancées méthodologiques et pratiques (possibilité de calculs) en évolution moléculaire/phylogénie

- Consensus (ici pas besoin d'arbre, à chaque site, aa ou nt le plus représenté) (très sensible à l'échantillonage)
- Parcimonie (bien pour des séquences proches)
- Maximum de vraissemblance (>1995)
- Analyse bayésienne

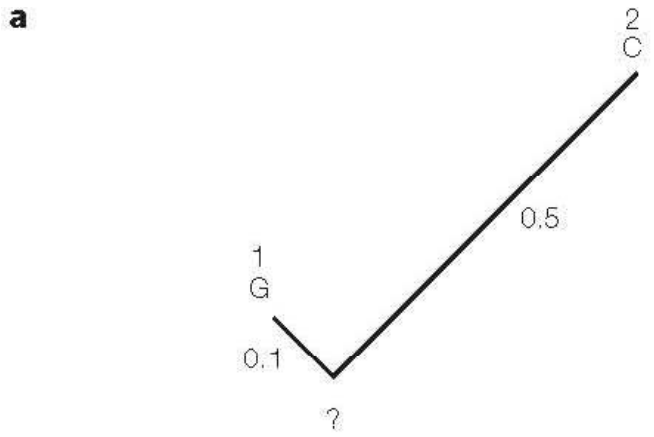
Parcimonie

Lorsque les séquences trop éloignées (possibilité accrue de subst. multiple) on obtient plusieurs reconstructions également parcimonieuses (ambiguité)

Possibilité d'appliquer certaines méthodes pour « choisir »

Même biais que pour la construction des arbres
-sous estimations des longueures de branches
-pas de modèle explicite d'évolution

Maximum Likelihood



Ancestral State (x)	Likelihood(x) = [Prob(1G) x Prob(2C)] | x	Posterior probability
C	0.031 x 0.635 = 0.020	0.14
G	**0.906 x 0.122 = 0.110**	**0.80**
A	0.031 x 0.122 = 0.004	0.03
T	0.031 x 0.122 = 0.004	0.03
Sum	0.138	1.00

On définit
-l'aligenemnt
-l'arbre (topologie, longueur de branche)
-le modèle (matrice de substitutions, variation du taux en fonction des sites)

On calcul la likelihood à chaque site pour chaque état de caractères (20 aa, 4 nt). On choisit l'état de caractère avec likelihood la meilleure.
On peut calculer un post. probabilité qui permet de déterminer la « robustesse »

Méthodes bayesiennes prennent en compte les incertitudes sur l'arbre, le modèle….

Exemple 1: Les RNases des ruminants

Premier article relatant une reconstruction ancestrale: 1990!

Pas d'autres possibilités que **consensus** et **parcimonie**

RNases impliquées dans croissance cellulaire

Récente évolution d'un rôle dans la digestion, peut-être chez les ruminants

But: tester la fonction d'une séquence ancestrale reconstruite

Buffle des marais

Buffle des rivières

Boeuf

Eland

5-10 Ma

Séquence ancestrale par parcimonie

Exemple 1: Les RNases des ruminants

Reconstruction non ambigüe aussi bien nucléotides que protéines

7 positions diffèrent entre les espèces dérivées

Toutes positions à la surface de la protéine sauf résidue 35

35=Met ou Leu

Structure 3D (cristal) suggère que changement de Leu par Met doit être associé à une autre mutation ponctuelle

Séquence actuelle en effet, si changement 35, **TOUJOURS** accompagné de changement 34 ou 37

OR séquence ancestrale=**PAS** de changement 34 ou 37!

QUE VA T'IL SE PASSER?

Exemple 1: Les RNases des ruminants

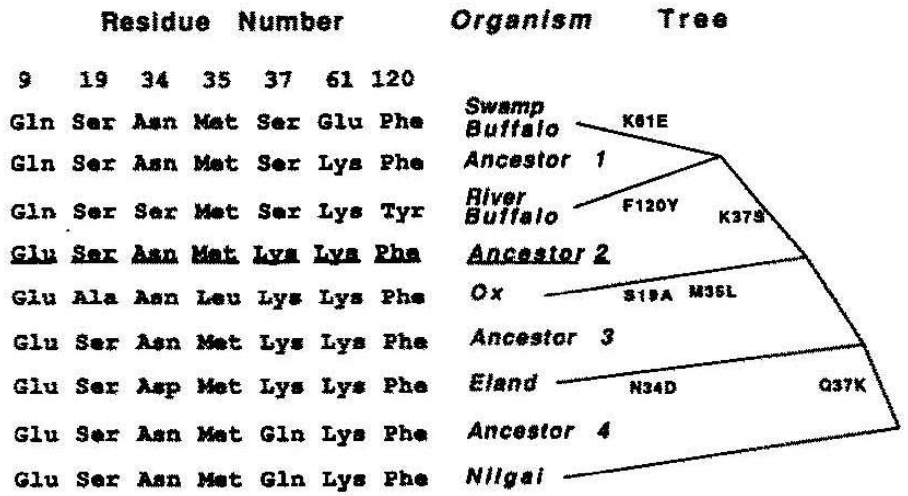


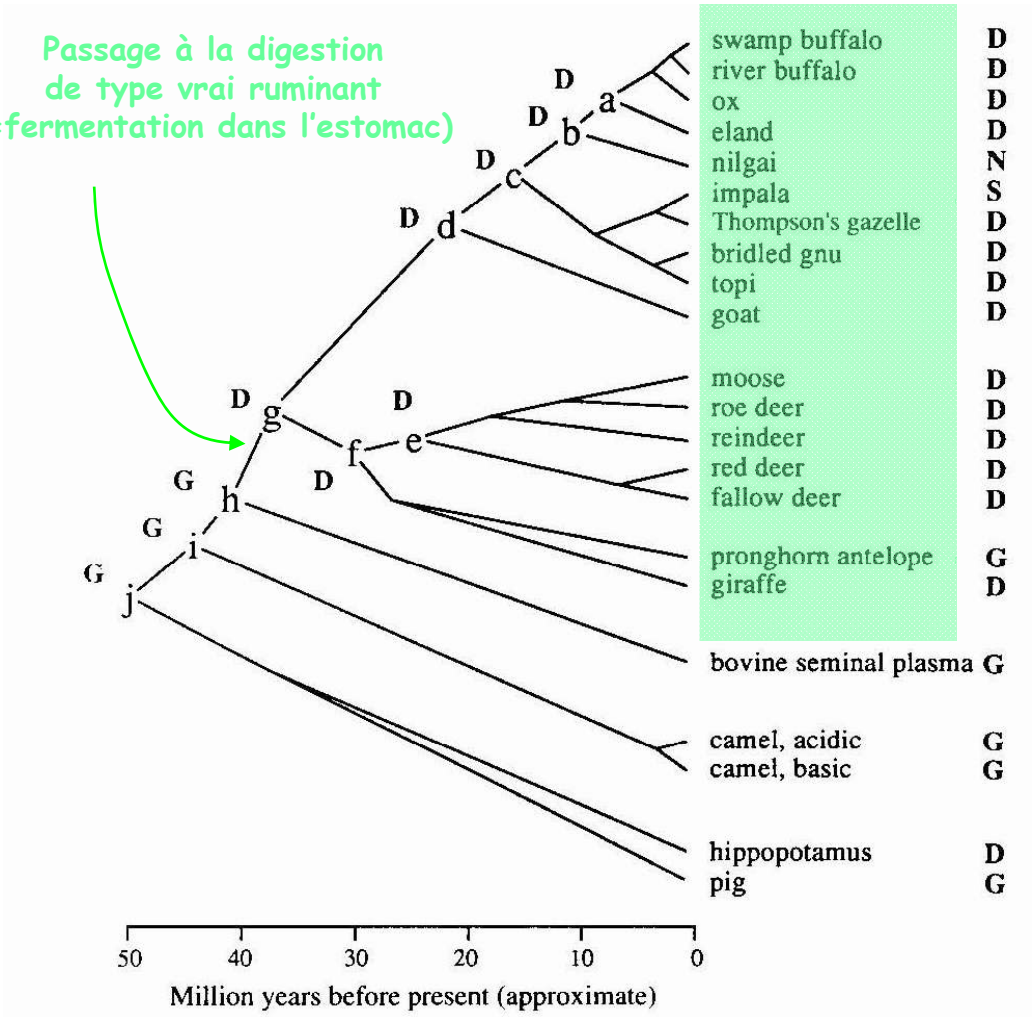
Fig.1. Sequences for ancient RNases reconstructed by parsimony. The reconstruction is based on a tree that relates the buffalos most closely (via Ancestor 1). The eland diverged prior to Ancestor 2. Unlisted amino acids are conserved within the first 3 organisms [2].

Reconstruction de la séquence ancestrale, production de la protéine, purification
Tests biochimiques:
La protéine fonctionne comme les RNAses actuelles!
Donc la reconstruction par parcimonie permet de recréer des protéines fonctionnelles

Attention, ici il n'y a que 2 changements d'AA/aux séquences actuelles!

Exemple 1: Les RNases 5 ans après

Les RNases digestives des artiodactyles



Reconstruction par parcimonie de tous les intermédiaires

Reconstruction de plusieurs séquences si ambiguité

Test (1) activité catalytique, (2) spécificité du substrat, (3) stabilité thermique

Activité catalytique contre petits ARNs et ARN simple brin OK

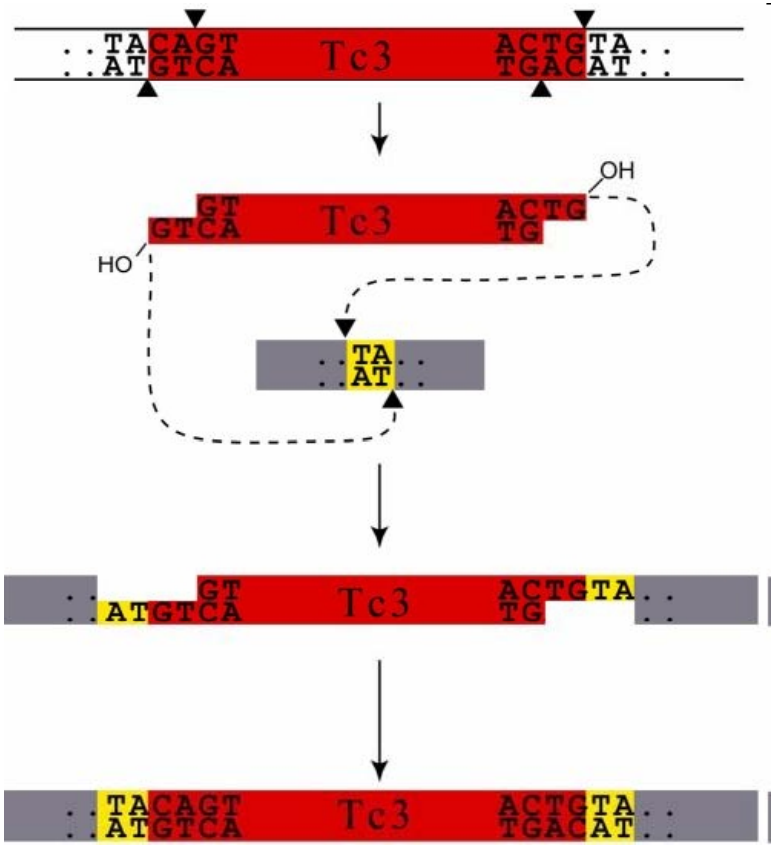
i et j thermostabilité décroît

activité catalytique 5X supérieure pour les duplex ARN pour h, i, j

Peut-être h,i,j non digestive

Rôle digestif apparaît chez les vrais ruminants

Exemple 2: Sleeping beauty



Exemple deTc3 de C.elegans

Famille Tc1/mariner DNA transposon

Existe des unicellulaires aux mammifères

Fonctionne avec un système type « cut-and-paste »

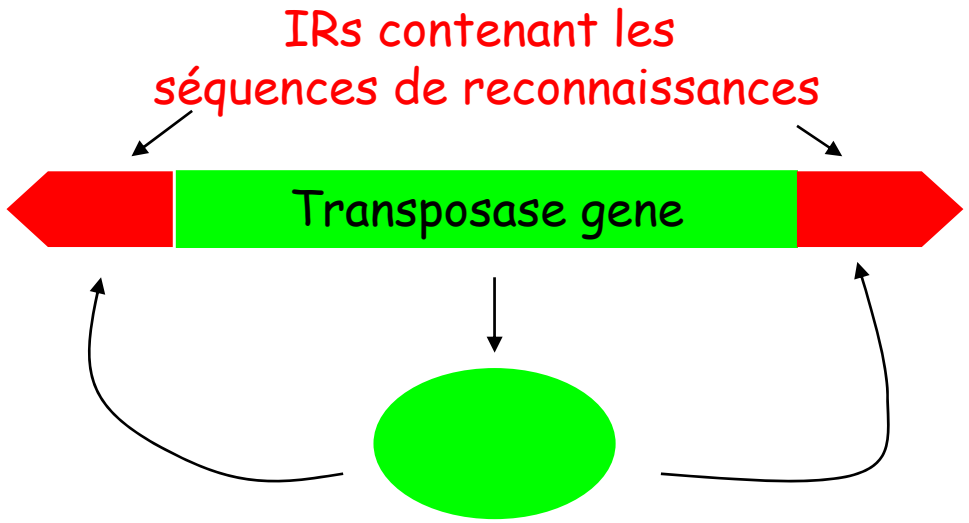
Pas de nécessité de facteurs spécifique de l'espèce

Transferts horizontaux mode d'infestation

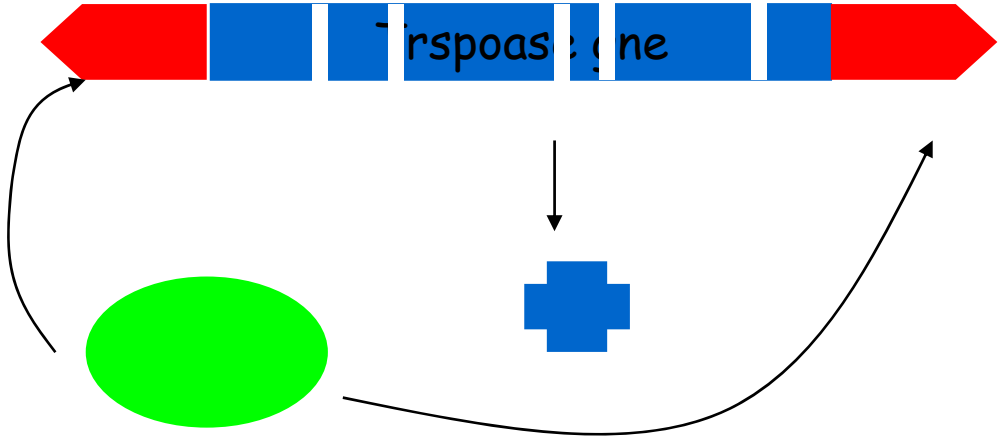
Exemple 2: Sleeping beauty

Transposons autonomes

Possèdent *cis* et *trans* composants



Transposons non-autonomes



Exemple 2: Sleeping beauty

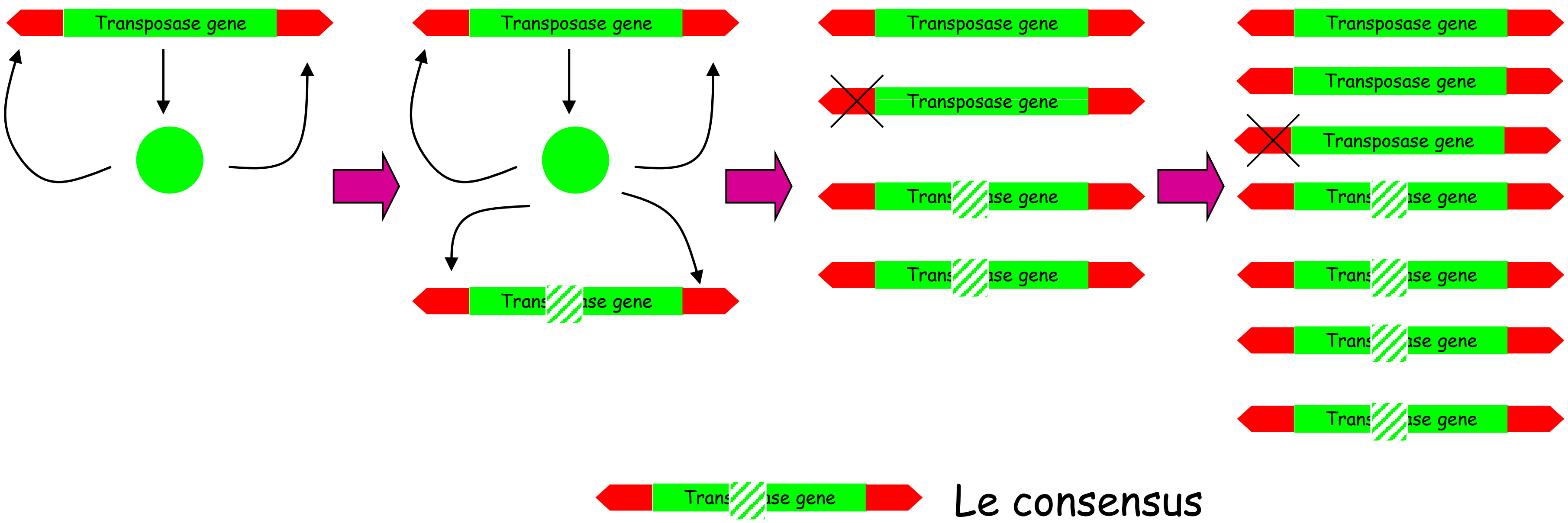
- Outils extrêmement puissant pour permettre l'insertion dans le génome d'un gène d'intérêt=transgenèse germinale, ou pour faire par exemple de l ' « enhancer trap » en utilisant un gène rapporteur (GFP)
- En 1997, pas de transposon ADN utilisable chez les vertébrés car aucun élément autonome connu
- Par contre, beaucoup de transposon de type Tc1/mariner non fonctionnels (inactifs) appelés TcE détectés dans plusieurs génomes de poissons
- Trois types de TcEs: type poisson-zèbre, salmonidés ou XenopeTXr. Salmonidés les plus jeunes et probablement les plus récemment actifs

Exemple 2: Sleeping beauty

Choix du mode de reconstruction: **Consensus**

Pourquoi? évolution non verticale mais horizontale, noeuds non déterminables de manière satisfaisante par parcimonie

Si consensus avec séquences d'une seule espèces: on obtient une séquence qui reflète la mutation à l'origine de l'inactivation du transposon dans cette espèce: **il faut les séquences chez plusieurs espèces**

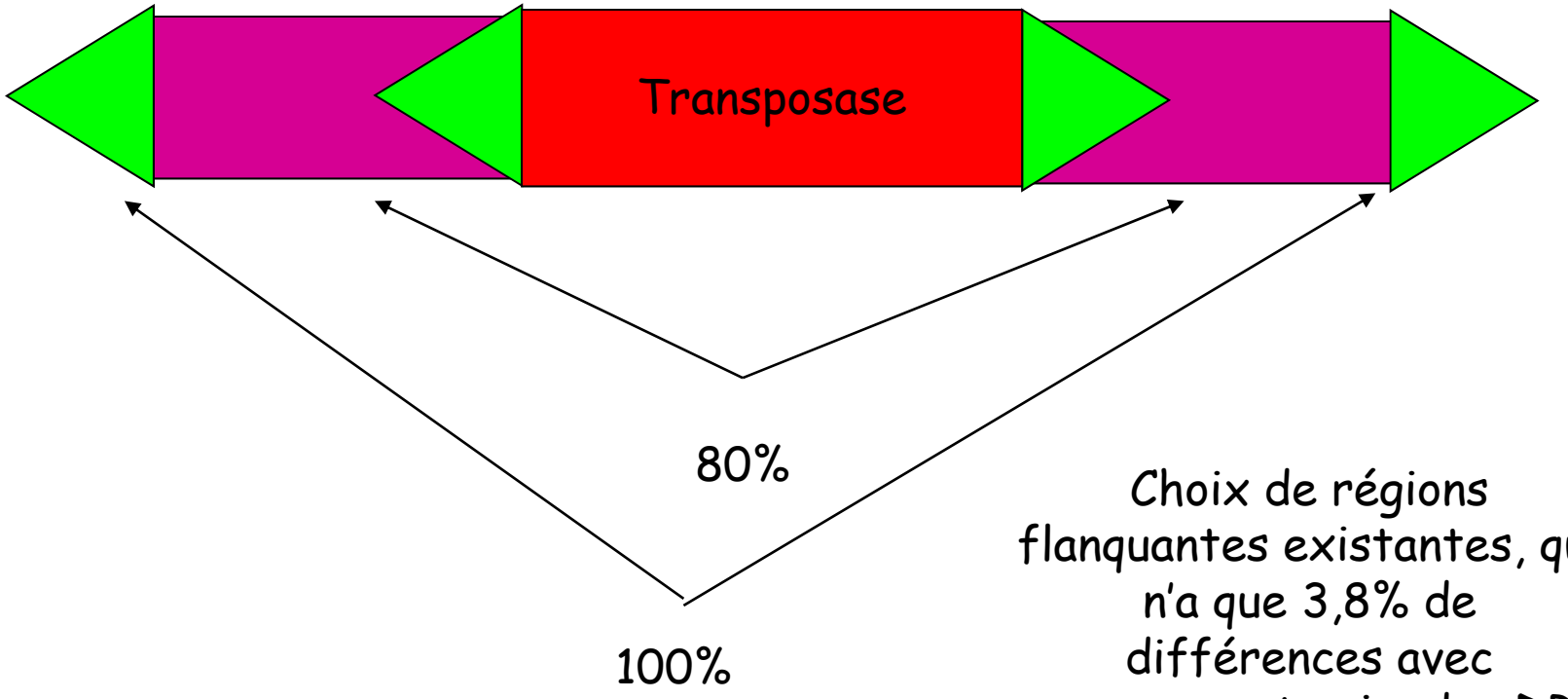


Exemple 2: Sleeping beauty

Transposase: Consensus majority-rule pour 12 séquences de TcE de type salmonidés échantillonés dans 8 espèces de poissons

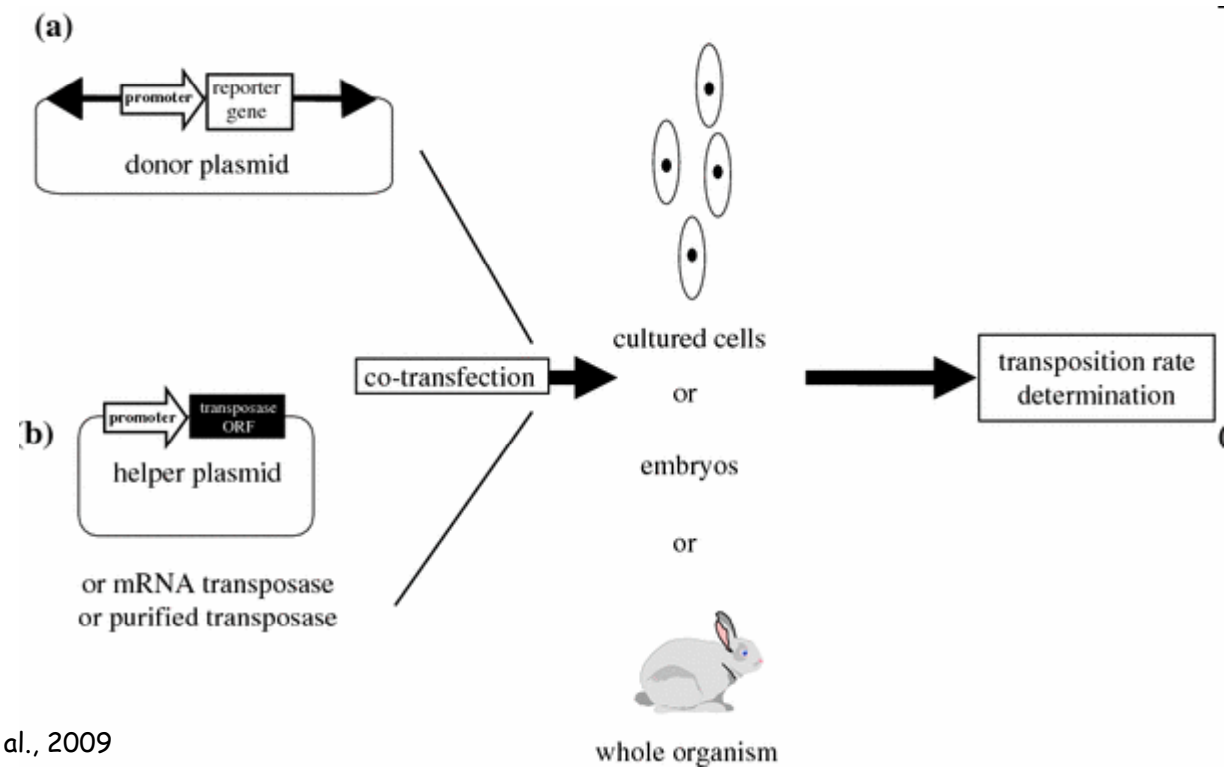
Distance: 10 millions d'années

IR/DR



Choix de régions flanquantes existantes, qui n'a que 3,8% de différences avec consensus et qui a des DRs intacts

Exemple 2: Sleeping beauty



Delaunière L et al., 2009

Ne mobilise pas les TcE internes, au moins chez les poissons (grande spécificité de reconnaissance de la séquences DR)

Exemple 3: Le Elongation Factor-Tu

1er article: 2003
Gaucher et al., 2003

Inférences contradictoires sur la température à laquelle vivaient les premières bactéries basées sur

GC content, température basse proposée par certains, distribution de la thermophilie chez les espèces actuelles.....proposition bactéries soit hyperthermophiles, soit mésophiles

L'ancêtre commun des bactéries actuelles était-il mésophile (20-40°C, 40-80°C, >80°C) thermophile ou hyperthermophiles??????????????????

EF-Tu: Protéine thermosensible et avec optimum à température physiologique (optimale de croissance), fixe GDP et régule la synthèse protéique

Reconstruction par ML du EF-Tu de l'ancêtre de toutes les Eubactéries

Saturation sites nucléiques travail en protéines

1 000 000 000 années !!!!

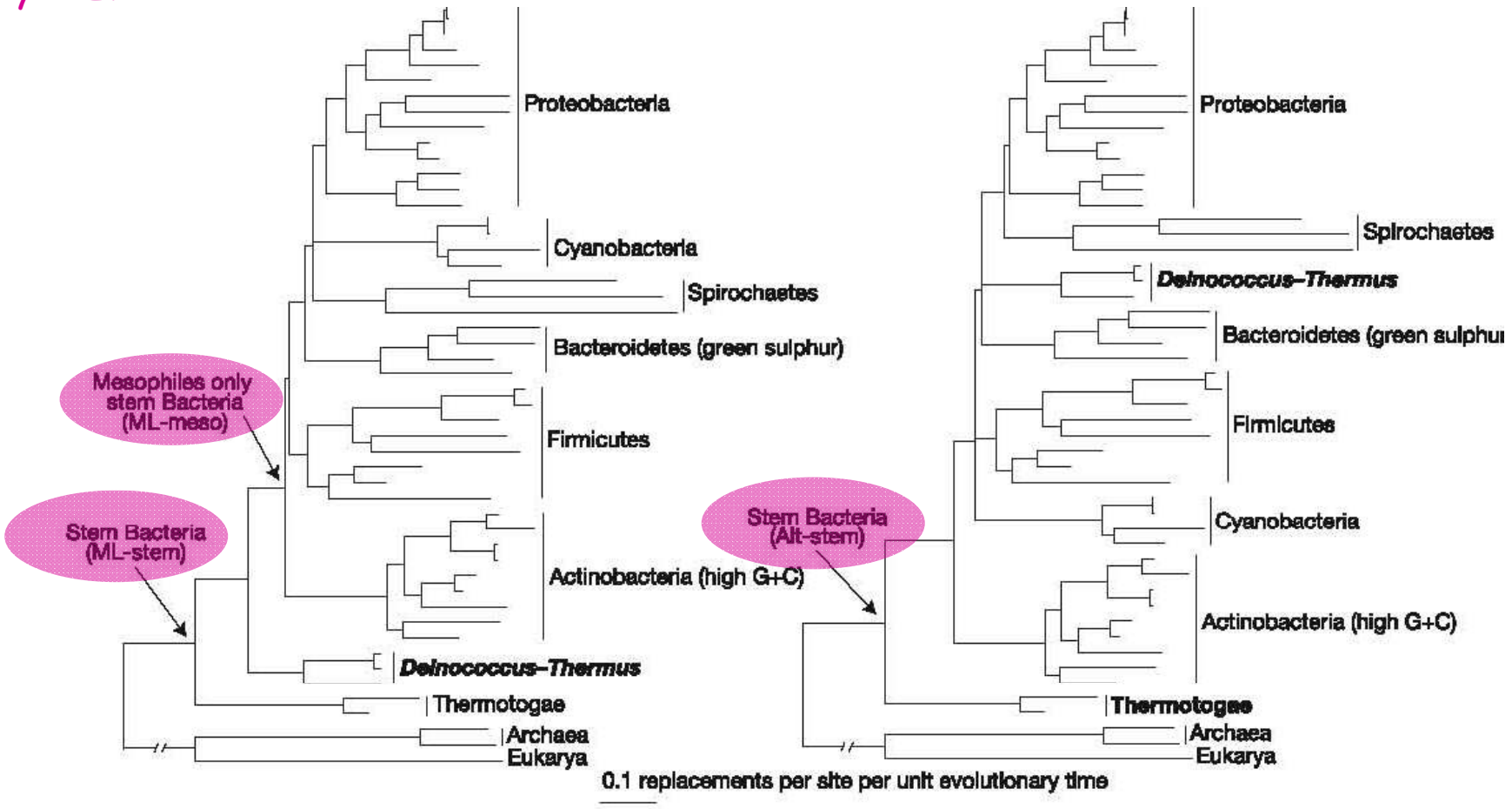
Exemple 3: Le Elongation Factor-Tu

ML gamma Prot

Each site, the *AA* avec la meilleure post proba

Phylo EF-Tu

Phylo Littérature



Exemple 3: Le Elongation Factor-Tu

- Problème d'une long branch attraction de *Aquifex*: espèce enlevée de la reconstruction
- Seq Archae et Eucaryote: entre 3-20: ne change rien
- Si on regarde les séquences, on peut penser (sans base soutenue) que ancêtre toutes bactéries étaient thermophile ou hyperthermophile et que l'ancêtre des mésophiles était mésophile.
- Tests *in vitro*: ML-stem et Alt-stem (seulement 93% d'identité de séquence) ont toutes deux un optimum à 65°C (pourtant Alt-stem est plus proche des séquences d'hyperthermophiles)
- ML-meso: optimum à 55°C (supérieur à l'optimum des mésophiles actuels)

Même s'il faut être prudents, ceci laisse supposé que l'ancêtre des bactéries actuelles était thermophiles!

Et pas de lien direct entre un %age d'identité de séquence et la thermosensibilité d'une protéine

Exemple 3: Le Elongation Factor-Tu

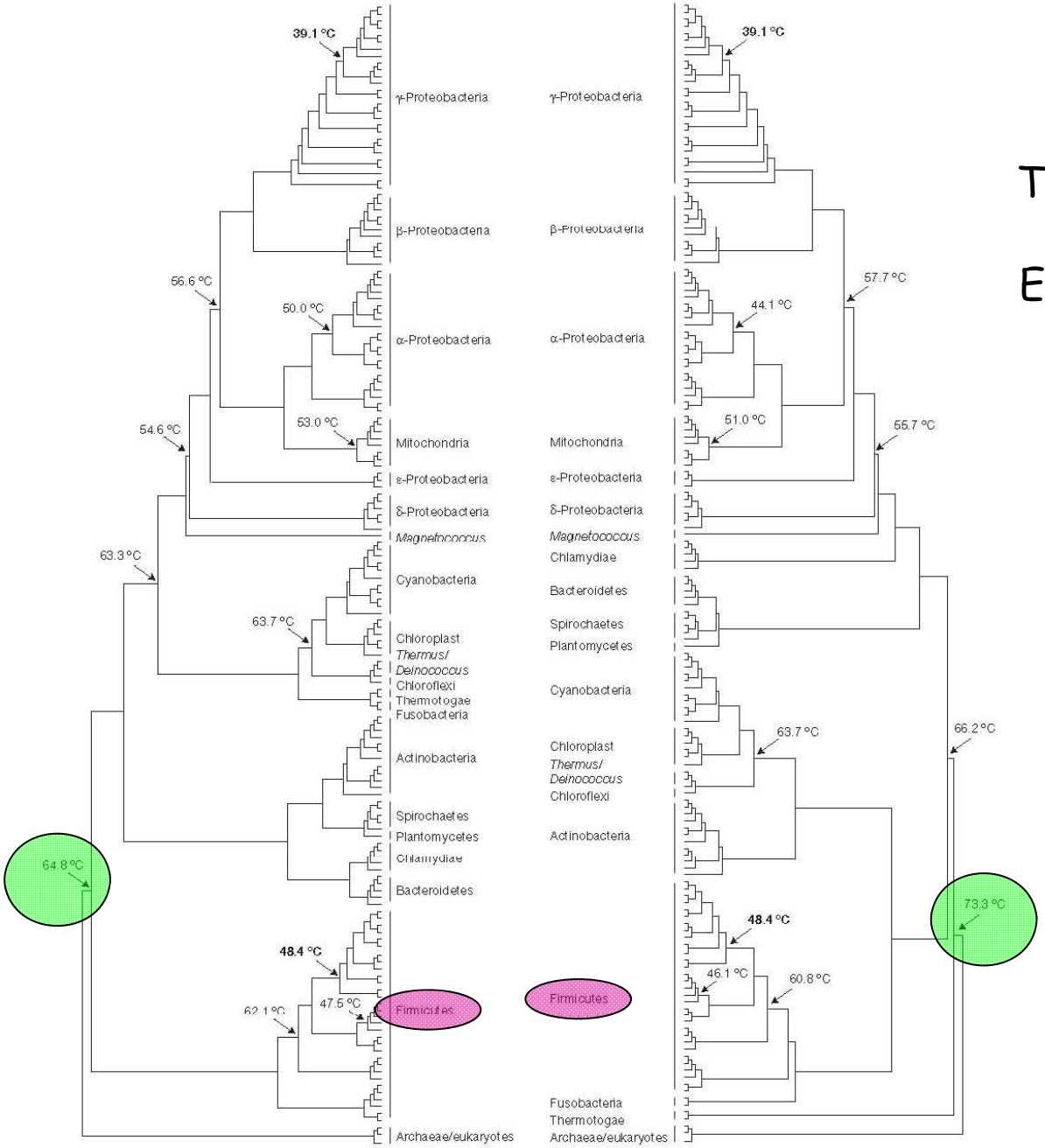
2ème article: Gaucher et al., 2008

La question est la même: à quelle température vivait l'ancêtre des eubactéries?

Cette fois-ci, test de plusieurs biais:

1er biais: la topologie de l'arbre: utilisation de deux topologies représentatives des deux vues classiques de relation entre les différentes bactéries (arbres basés sur un grand nombre de séquence)

Test de la Tm des ancêtre reconstruits à partir des deux topologies



Tendance identique dans les deux cas :
EF les plus ancestrales sont les plus thermostables, noeuds intermédiaire les moins thermostables

Calcul que la similarité des Tm (9°C) est significative sachant que les deux prot n'ont que 78% d'identité de séquence

(quel test?)

Exemple 3: Le Elongation Factor-Tu

2ème biais: la matrice de substitution utilisée a été construite à partir de séquence actuelles, cela ne représente peut-être pas la matrice de substitution ancestrale

exemple: les AA hydrophobes ont une valeur élevée de fréquence d'équilibre dans les matrices de substitution: on peut avoir un biais vers une séquence ancestrale riche en AA hydrophobes

Récupération des fréquences des différents AA d'ancêtres de 31 familles de gènes reconstruits à partir de 16 espèces actuelles (8 mésophiles et 8 thermophiles) et utilisation de ces fréquences comme fréquences d'équilibre de la matrice de substitution.

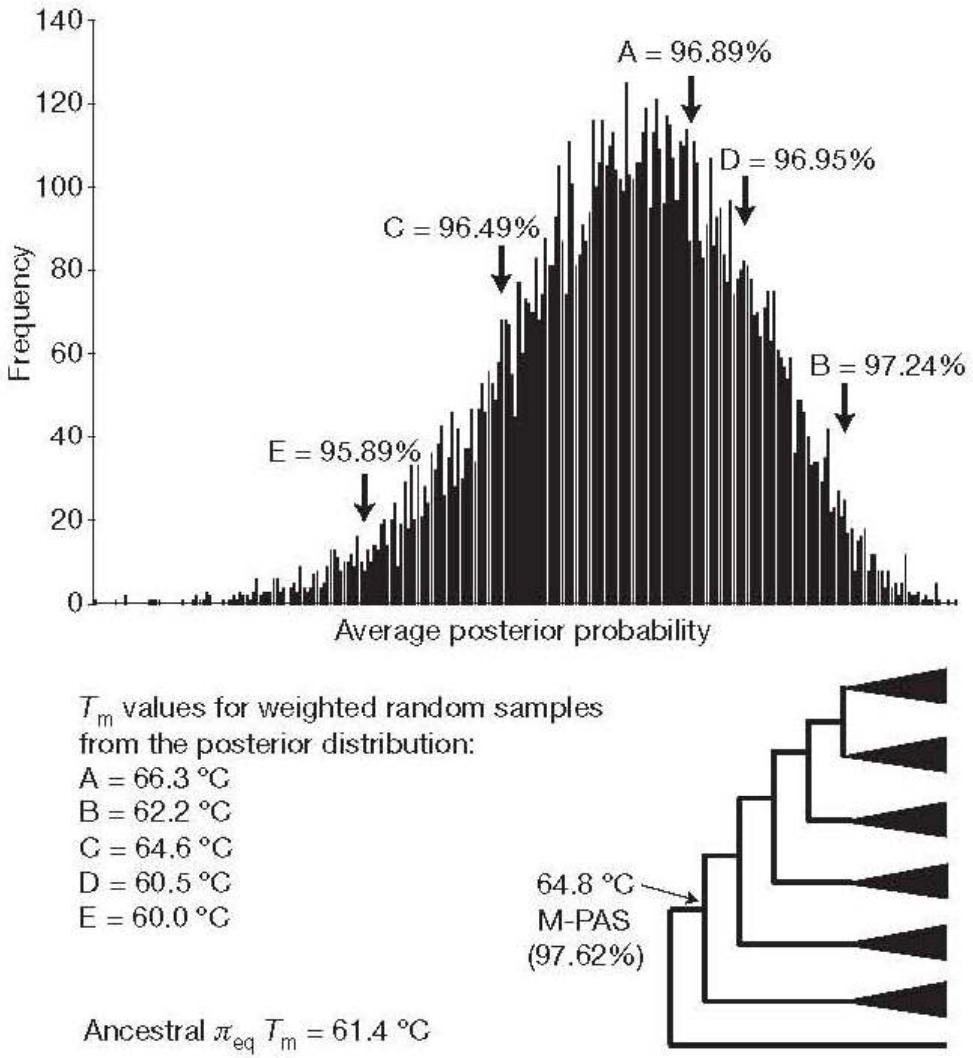
Tout cela pour reconstruire la séquence ancestrale en utilisant ces données et la topologie du premier arbre: Tm=61,4°C/ Tm=64,8°C

Exemple 3: Le Elongation Factor-Tu

3ème biais: La séquence ancestrale choisie correspond à la séquence ou à chaque site on choisit l'aa avec la meilleure probabilité postérieure= M-PAS (Most Probabilistic Ancestral Sequence)

Cela implique que chaque site est indépendant des autres, ce qui dans l'absolu n'est pas vrai

Exemple 3: Le Elongation Factor-Tu



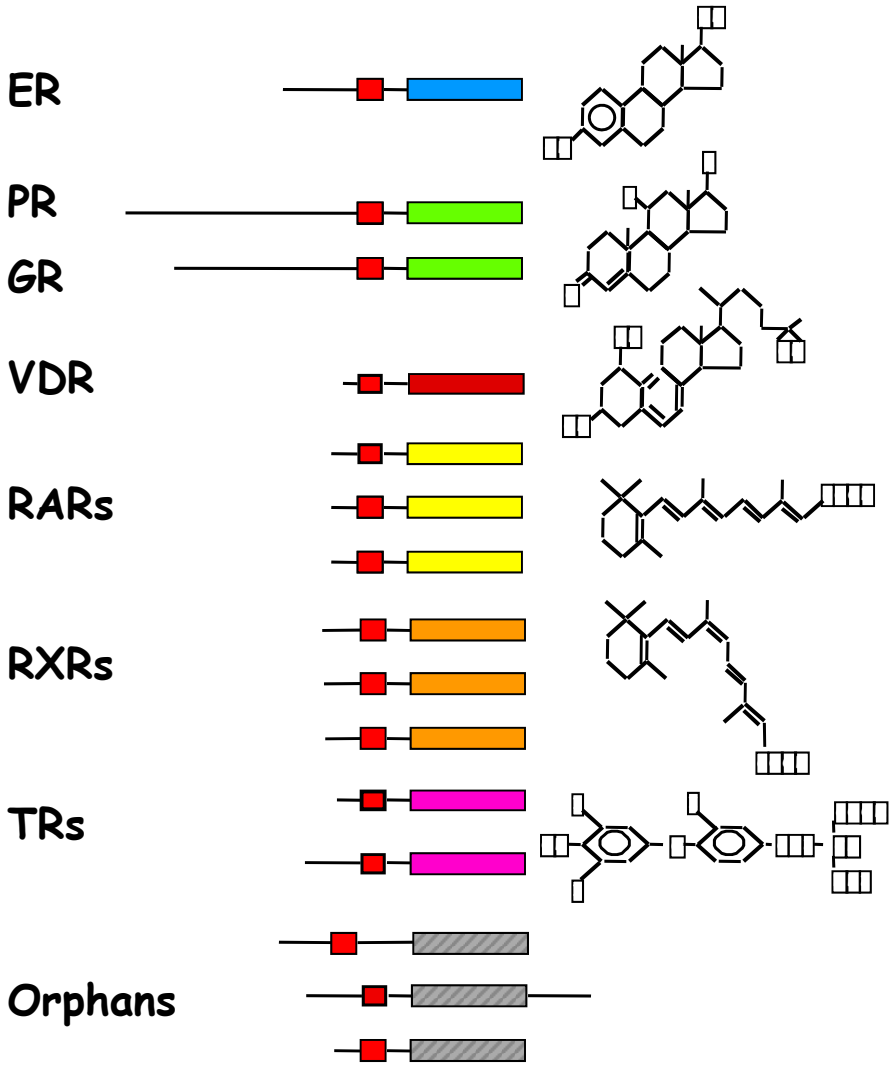
Sampling de 10 000 Seq ancestrales

5 séquences reconstruites au hasard avec des post-proba globales différentes

Test Tm

Résultats consistants avec un ancêtre thermophile!

Exemple 4: Les récepteurs nucléaires



Oestrogènes

Progestérone

Vitamine D

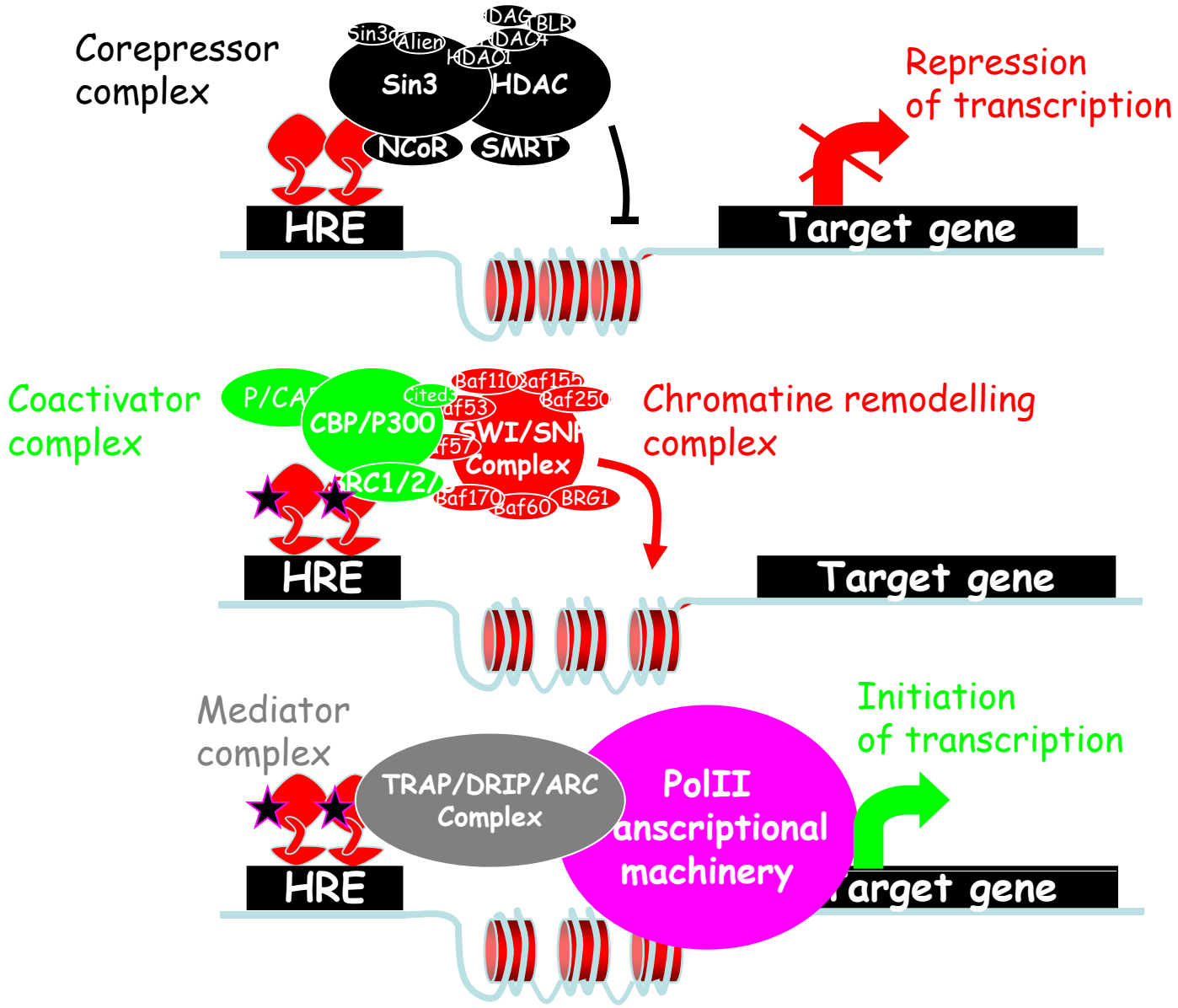
Acides rétinoïques

Hormone thyroïdienne

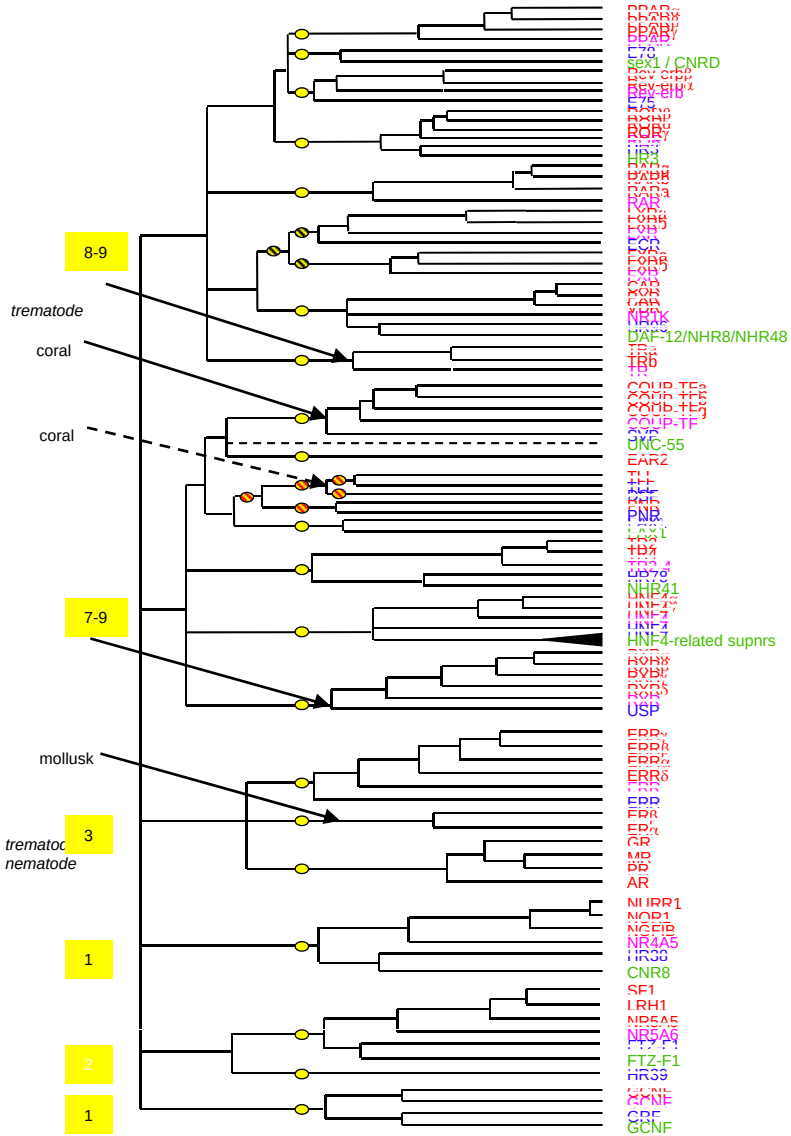
?

- **Domaines bien conservés**
- **Grande variété de ligands**
- **Grande variété d'action**

Exemple 4: Les récepteurs nucléaires



Exemple 4: Les récepteurs nucléaires

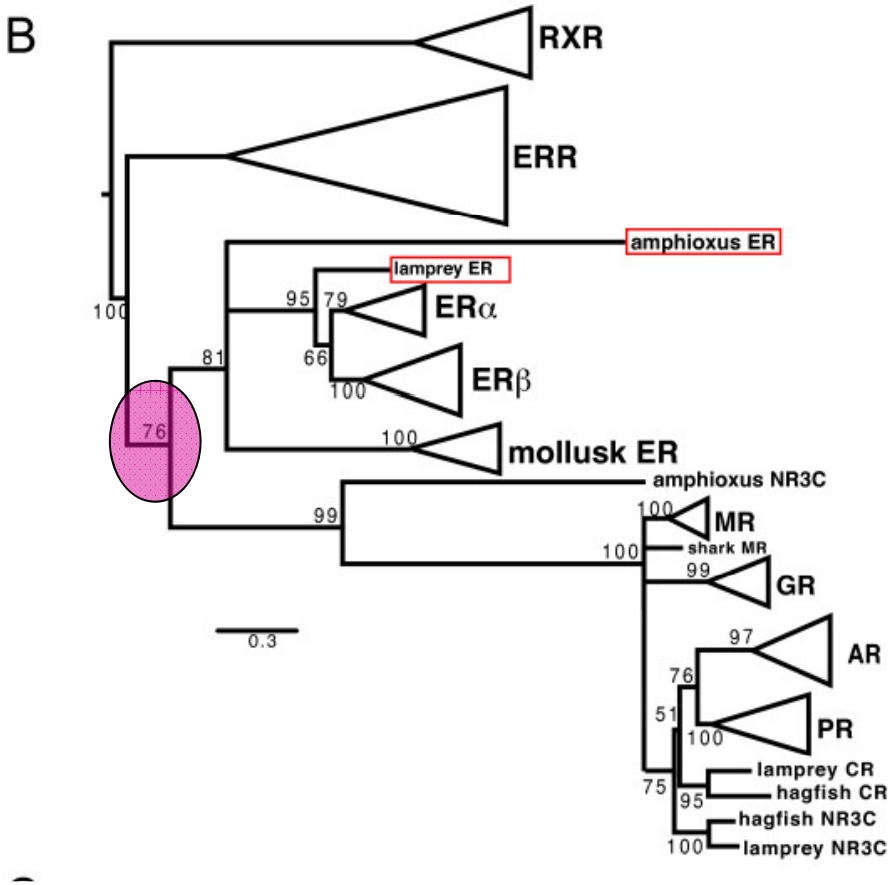


- 6 (7, 8?) sous-familles, déjà présentes chez les premiers métazoaires
- Diversifications des sous-familles avant Urbilateria
- Duplications chez vertébrés

Sous-famille NR3: stéroïdes

Sous-famille NR5: orphelins?

Exemple 4: Les récepteurs nucléaires

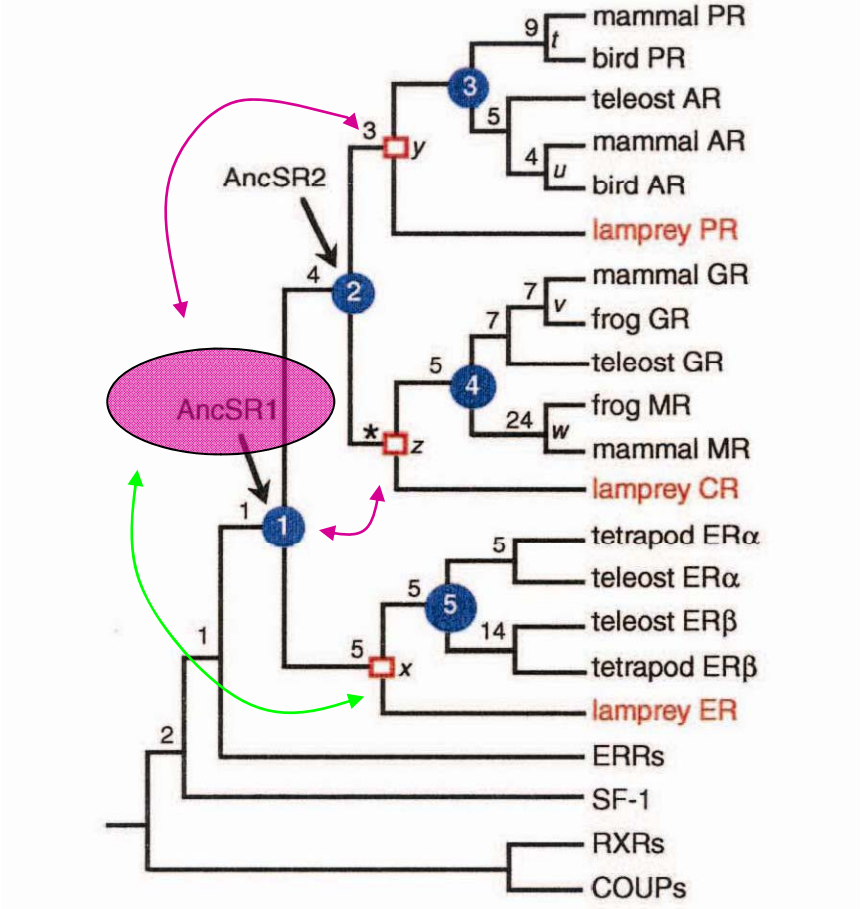


La sous-famille 3: ERRs, ERs et SRs

ER et SR longtemps connus uniquement chez vertébrés

Evolution récepteurs/ligands: quel ligand fixait le récepteur ancêtre de ERs et SRs?

Exemple 4: Les récepteurs nucléaires



2001: pas de ER ou SR en dehors des vertébrés

Phylogénie par parcimonie

Relative Rate Test (est-ce que deux séquences ont le même taux d'évolution?)

rate of AA replacement 4 fois plus élévé dans la lignée menant à y, z que dans celle menant à x

Reconstruction « *in silico* »

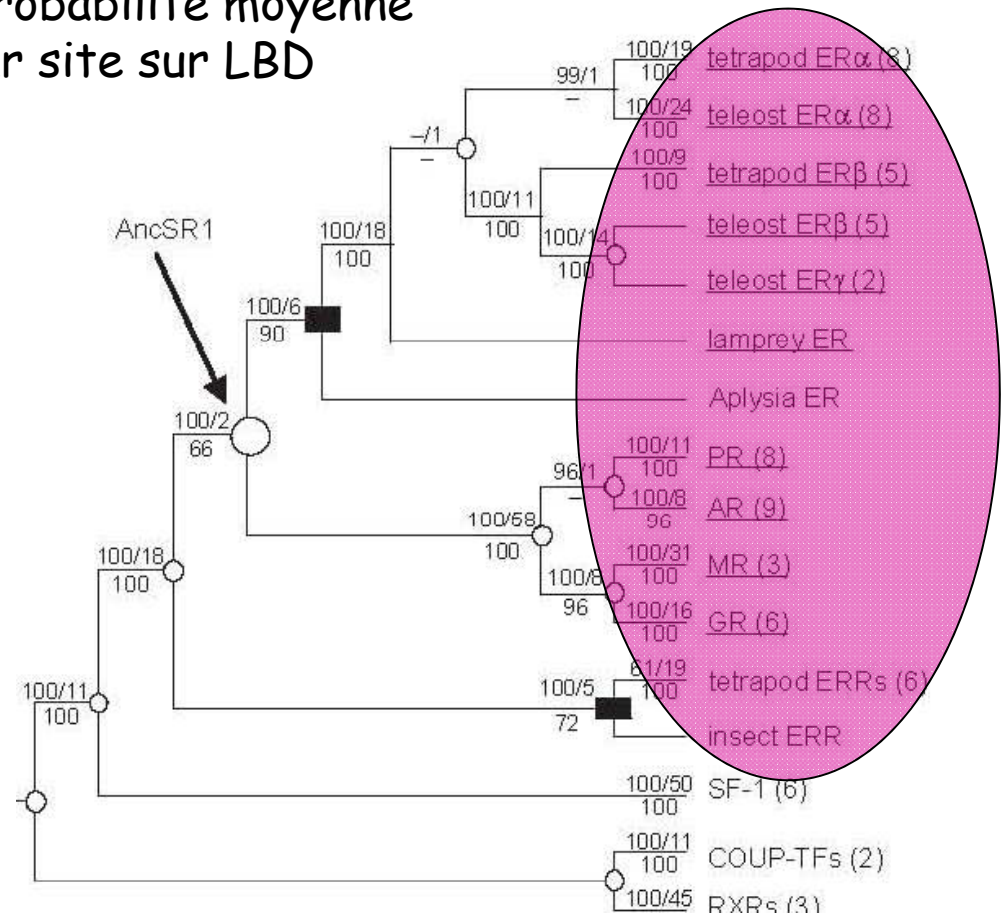
Analyse des caractéristiques de la séquence (positions de contact avec ligands....)

Conclusion (abusive?): AncSR1 était un ER

Exemple 4: Les récepteurs nucléaires

2003: un ER chez un mollusque!

62% probabilité moyenne par site sur LBD



Sous-famille 3

Tests sur ER Aplysia:

Fixe ERE

Est un activateur constitutif

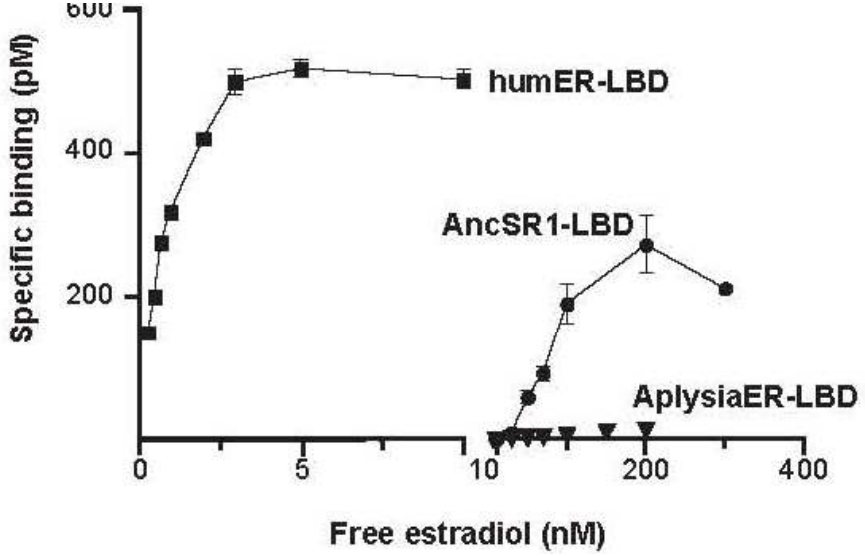
Exemple 4: Les récepteurs nucléaires

Reconstruction de AncSR1, test des DBD et LBD fusionnés

DNA binding sur ERE

LBD transactivation avec estradiol (+ faible que humains ERa et ERb)

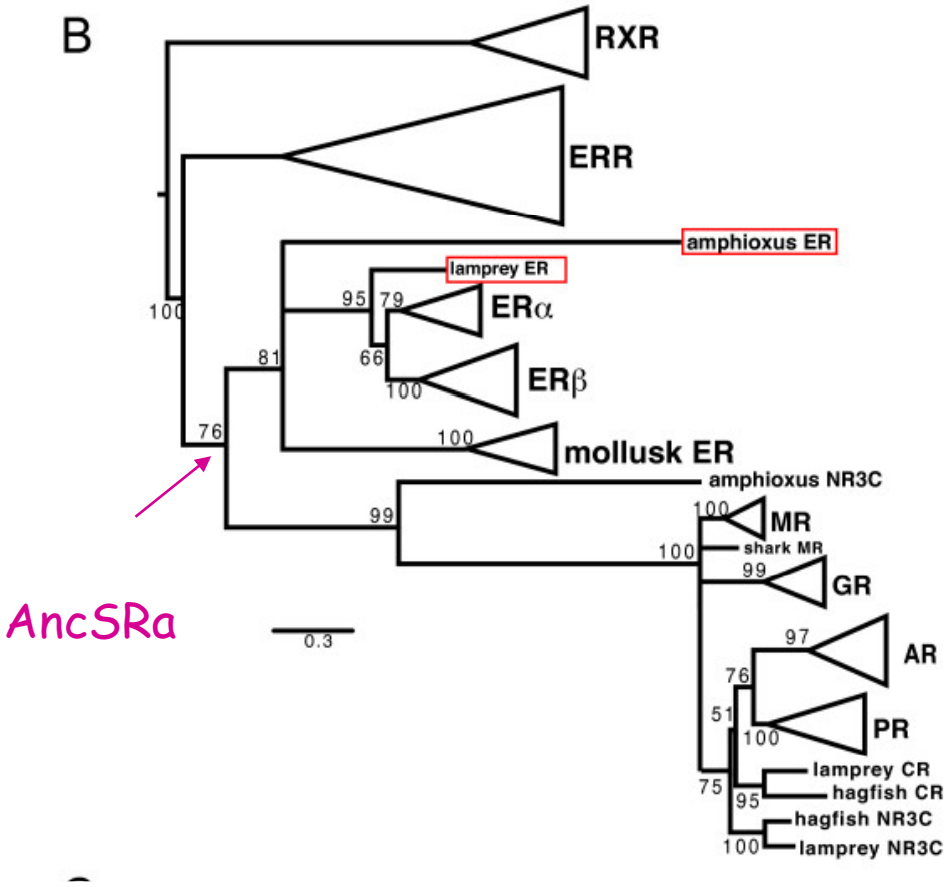
LBD aussi transactivation avec d'autres stéroïdes mais + faible



Hormone	EC_{50} (nM)	RAE
Estriol	323	1.40
Estradiol	37	1.00
Estrone	416	0.58
Progesterone	>1,000,000	0.45
Testosterone	3981	0.37
Dihydrotestosterone	9772	0.17
Androstenedione	-	0.01
Corticosterone	-	0.04
Cortisol	-	0.03

Max fold increase X/max fold increase estradiol

Exemple 4: Les récepteurs nucléaires



2008: plus de séquences

Un ER d'amphioxus

Un SR d'amphioxus

ER ne fixe pas l'estradiol!

Reconstruction de AncSRa

Positions connues pour empêcher la liaison à estradiol quand mutées chez homme ressemble à AA chez les mollusques

Si l'on refait la phylogénie avec AncSR1, il se place avec les ERs!

L'échantillonage taxonomique est déterminant!

Plusieurs biais à prendre en compte

Pour résumer:

- L'échantillonage taxonomique
- La topologie bien évidemment
- Le modèle d'évolution
- La reconstruction en elle-même en utilisant à chaque site le caractère avec la probabilité postérieure la plus importante (supposition de l'indépendance des caractères entre eux)

Lorsque l'on prend en compte toutes ces incertitudes, il est cependant possible de tirer des conslusions extrêmement intéressantes impossibles à obtenir par une autre méthode

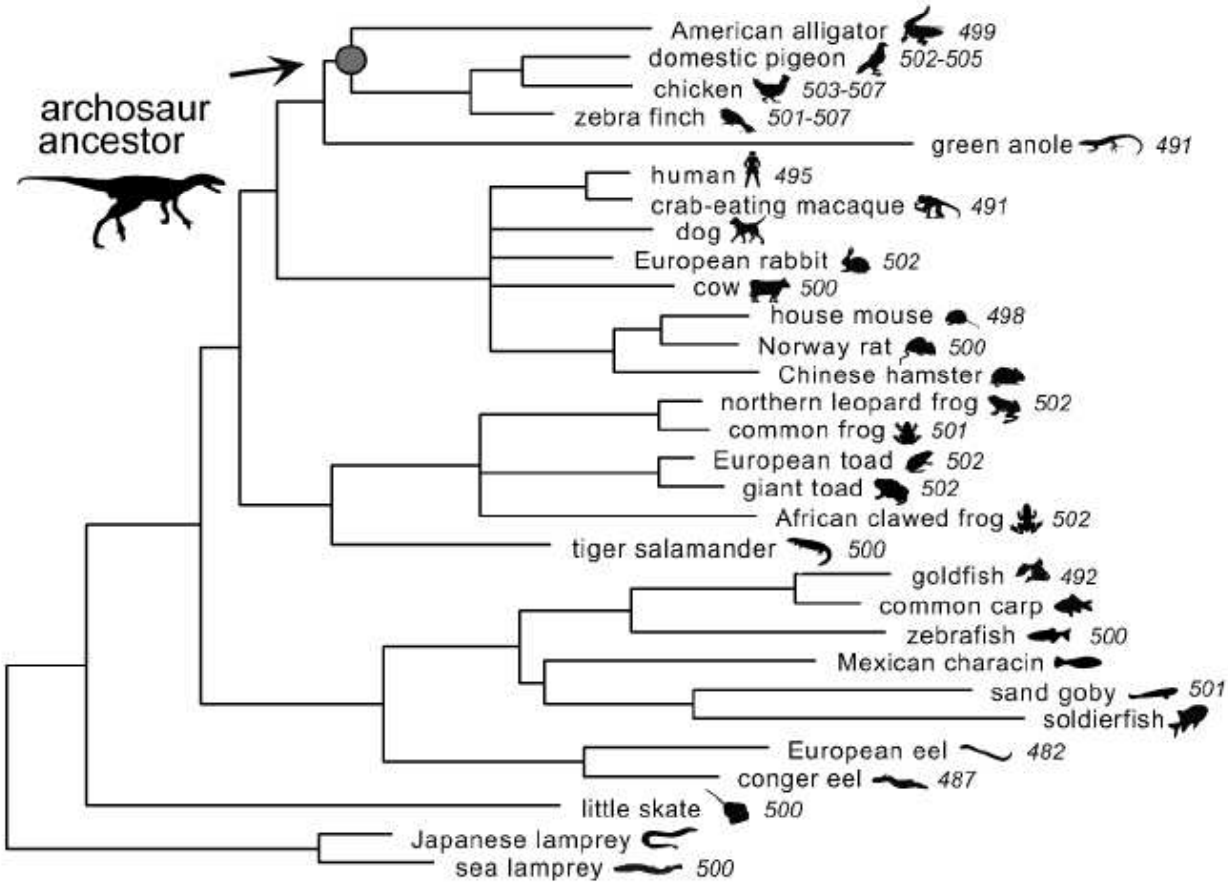
Visual pigment

Pigment intervenant dans la vision: rhodopsin

Archosaur regroupe crocodiles, oiseaux: quel était le mode de vie de l'ancêtre?

La rhodopsin ancestral absorbe à
508nm, ce qui est « plus rouge » que chez n'importe quel autre vertébré actuel

Résultat qui supporte l'hypopothèse d'un ancêtre ayant un mode de vie plutôt nocturne



Chang et al., 2002