

## Evolution Moléculaire et Bioinformatique

### Projet **MODEL\_PHYLO**

UMR 517I GPIA  
Montpellier

**N. Galtier**  
K. Belkhir  
J. Dutheil

UMR 5506 LIRM  
Montpellier

O. Gascuel  
N. Lartillot  
S. Blanquart

UMR 5558 BBE  
Lyon

M. Gouy  
G. Perrière  
B. Boussau

University College  
London

Z. Yang

## **L'évolution moléculaire:**

- née dans les années 70 de la confrontation de la **théorie synthétique de l'évolution** aux premières **données moléculaires**.
- construite autour des **conflits** neutralisme/sélectionnisme, cladistique/statistique
- période **pionnière** au cours de laquelle les questions fondamentales ont été posées.

## **La génomique évolutive:**

- les génomes complets ont modifié l'**ordre de grandeur** des jeux de données
- transcriptomique, protéomique, réseaux métaboliques: vers une **biologie intégrative**
- nécessité de l'outil **bioinformatique**

## Sept bonnes raisons pour faire de la génomique évolutive

**annoter** les génomes (→ génomique fonctionnelle)

comprendre les **mécanismes** de l'évolution moléculaire (→ génomique structurale)

comprendre la mise en place des **plans d'organisation** (→ évo-dévo)

caractériser les gènes de **l'adaptation** (→ amélioration des espèces domestiquées)

mesurer/appréhender la **biodiversité** (→ gestion, conservation des écosystèmes)

reconstruire **l'histoire** des espèces (→ paléontologie)

caractériser la dynamique des **interactions durables** (→ épidémiologie, microbiologie, virologie)

## **Génomique Evolutive et Bioinformatique**

### **Une discipline intrinsèquement demandeuse de bioinformatique:**

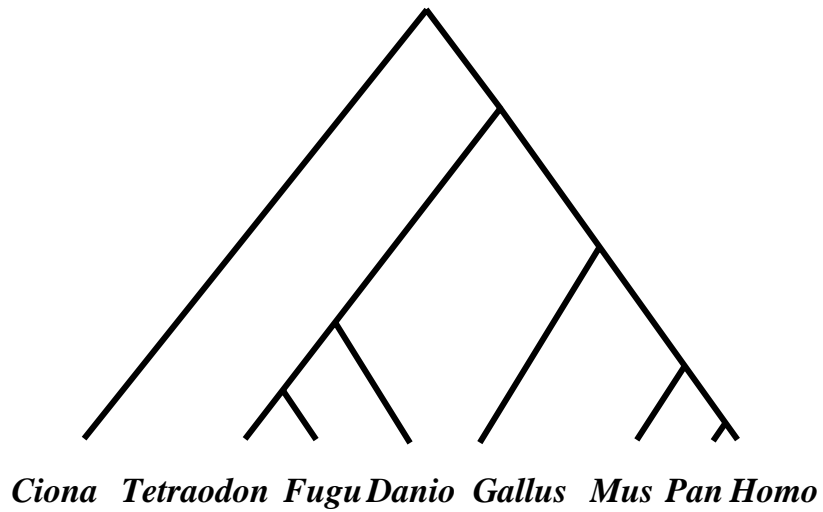
- le volume des données est à croissance exponentielle
- reconstruire le passé nécessite de modéliser

### **Un contact aisé entre biologistes et informaticiens:**

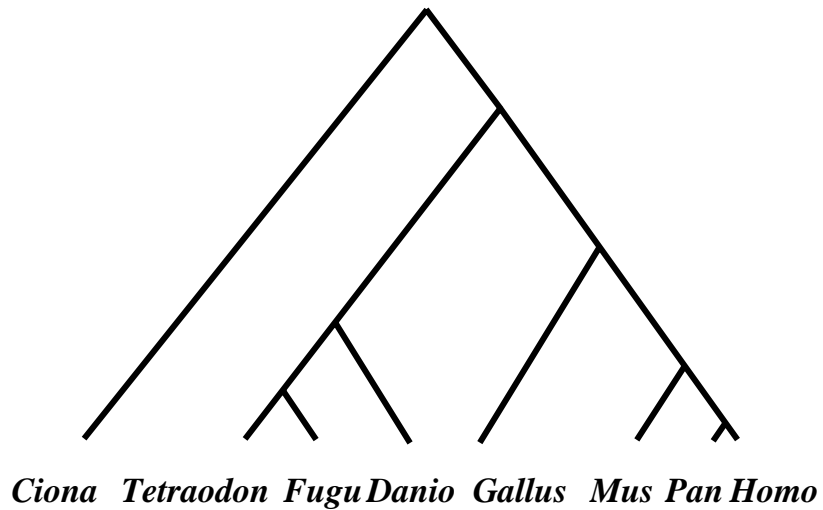
- tradition de biologistes théoriciens en génétique évolutive
- tradition de biostatisticiens pour appréhender la biodiversité
- les arbres (graphes), les séquences (mots), les processus (chaînes de Markov) et leurs propriétés font partie de la culture générale des informaticiens.

**→ un des piliers de la bioinformatique, en France (Alphy) et internationalement (MEP)**

## Les phylogénies, outils de base de la biologie évolutive

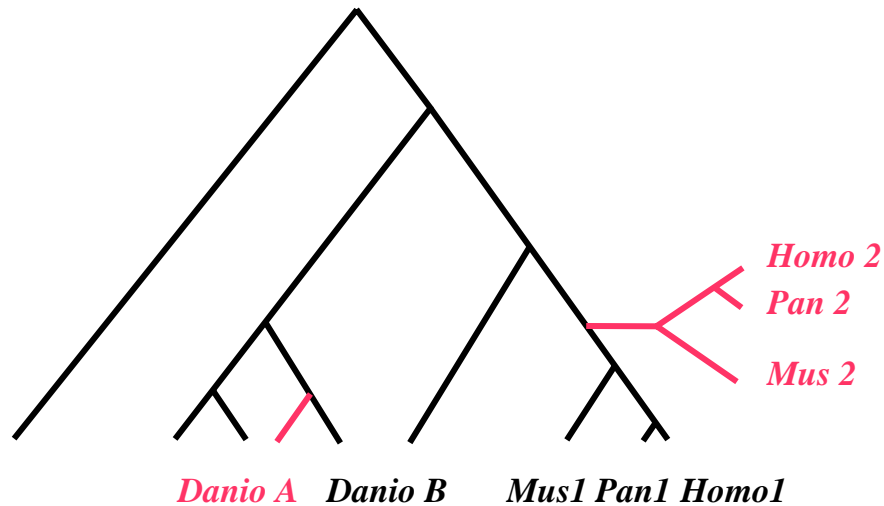


## Les phylogénies, outils de base de la biologie évolutive



- systématique

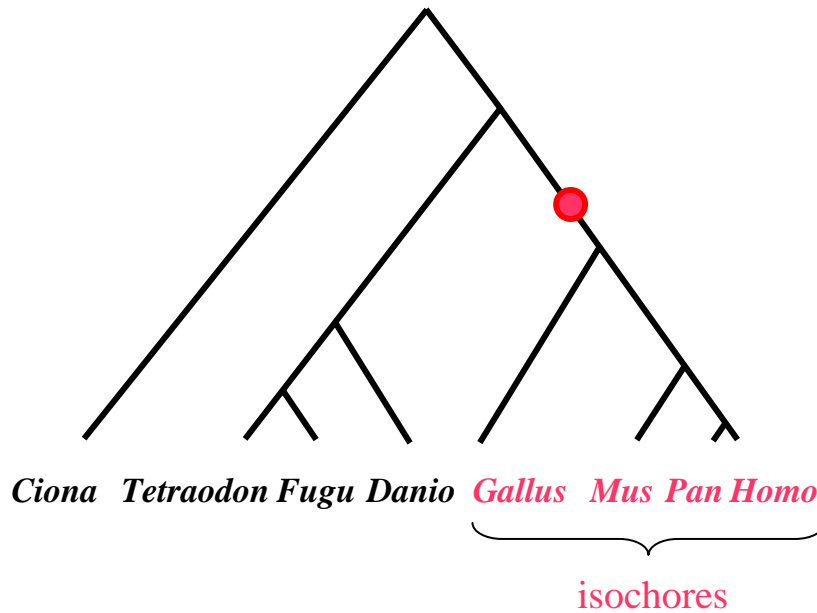
## Les phylogénies, outils de base de la biologie évolutive



- systématique

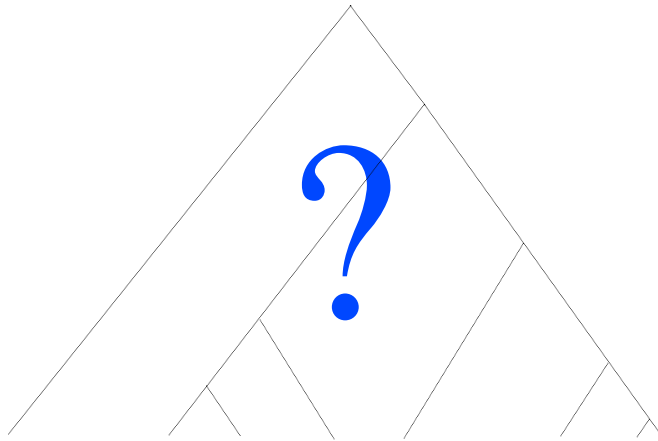
- annotation, évo-dévo

## Les phylogénies, outils de base de la biologie évolutive



- systématique
- annotation, évo-dévo
- **génomique comparative**

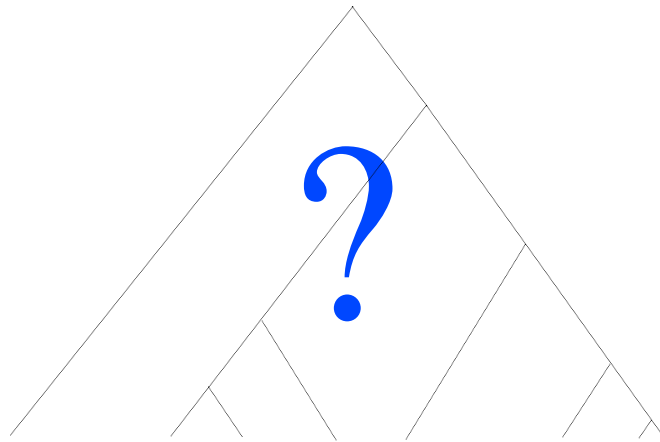
## Les phylogénies, outils de base de la biologie évolutive



*Ciona Tetraodon Fugu Danio Gallus Mus Pan Homo*

## Les phylogénies, outils de base de la biologie évolutive

...ATGACAGT...



*Ciona* *Tetraodon* *Fugu* *Danio* *Gallus* *Mus* *Pan* *Homo*

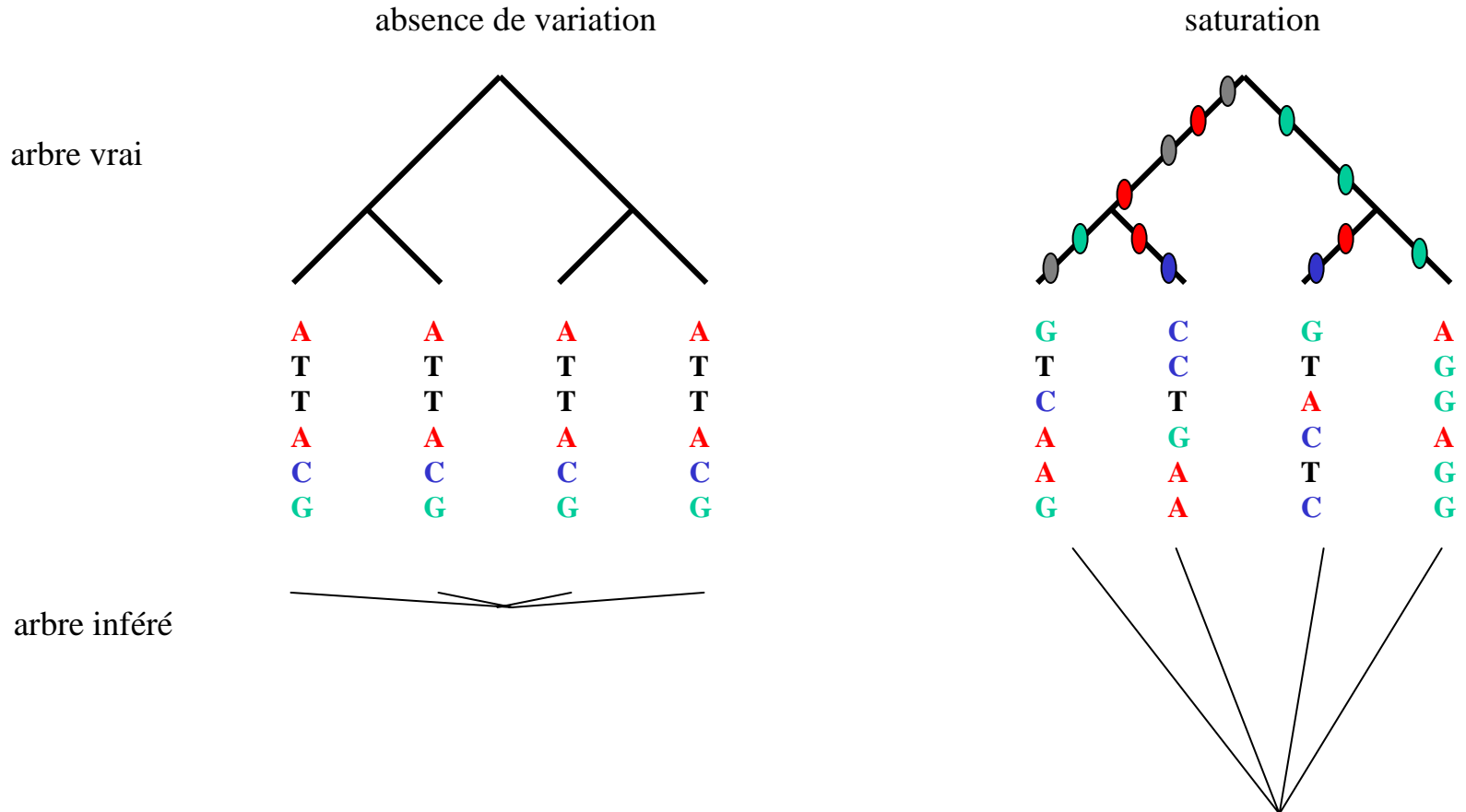
G	A	A	A	A	A	A	A
T	T	G	T	T	C	C	C
C	T	T	G	G	G	G	G
C	A	A	A	C	C	C	T
A	C	C	C	C	C	C	C
A	G	A	A	A	A	A	A
A	G	G	G	G	G	G	G
T	C	C	C	A	A	A	A

L'objectif de la phylogénie moléculaire:

- reconstruire l'**arbre**
- estimer les **longueurs de branches**  
(taux d'évolution, dates de divergence)
- caractériser le **processus** évolutif
- reconstruire les **génomés ancestraux**

## Phylogénomique: les défis méthodologiques

- un rapport **signal / bruit** incertain



## Phylogénomique: les défis méthodologiques

- un rapport **signal / bruit** incertain
  
- des processus évolutifs **complexes**
  - écarts à l'**horloge moléculaire**
  - **variation de vitesse** entre sites
  - **hétérotachie** = covariations
  - **coévolution** entre sites
  - **fréquences stationnaires** variables entre sites et lignées
  - **transferts latéraux**

## Phylogénomique: les défis méthodologiques

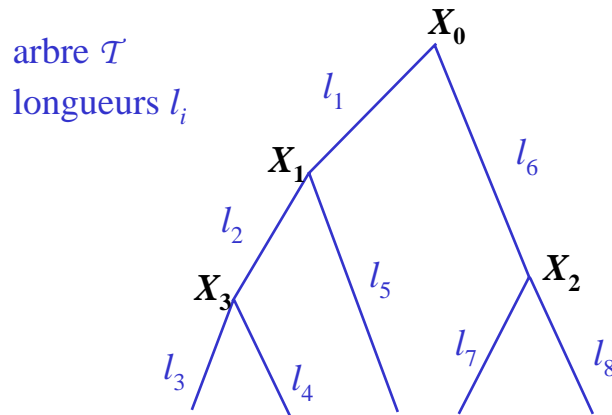
- un rapport **signal / bruit** incertain
- des processus évolutifs **complexes**
- des données **hétérogènes**

	gène A	gène B	gène C	gène D	gène E
espèce 1		————	————		————
espèce 2	————	————	————	———	————
espèce 3	————	————	————	———	————
espèce 4		————	————	———	
espèce 5		————		———	————
espèce 6		————			————
espèce 7	————	————			————
espèce 8	————	————			————

## Phylogénomique: les défis méthodologiques

- un rapport **signal / bruit** incertain
- des processus évolutifs **complexes**
- des données **hétérogènes**
- des analyses **coûteuses** en temps de calcul

## Modélisation statistique en phylogénie moléculaire



générateur Markovien :  $\mathbf{M}$

	A	C	G	T
A		$\beta$	$\alpha$	$\beta$
C	$\beta$		$\beta$	$\alpha$
G	$\alpha$	$\beta$		$\beta$
T	$\beta$	$\alpha$	$\beta$	

données :  $\mathbf{D}$

$d_1$ :	A	A	C	A	G
$d_2$ :	T	T	C	T	T
$d_3$ :	A	A	A	A	A

$$L(l_i, \mathbf{M}, \mathcal{T}) = \Pr(\mathbf{D} \mid l_i, \mathbf{M}, \mathcal{T}) = \prod_i \Pr(d_i \mid l_i, \mathbf{M}, \mathcal{T})$$

Calcul de vraisemblance:

$$\Pr(d_i) = \sum_{x_0} \Pr(X_0=x_0) \Pr(d_i \mid x_0)$$

$$\Pr(d_i \mid x_0) = \sum_{x_1} \sum_{x_2} \Pr(x_0 \rightarrow x_1, l_1) \Pr(d_i \mid x_1) \Pr(x_0 \rightarrow x_2, l_2) \Pr(d_i \mid x_2)$$

## Maximum de vraisemblance et optimisation

- estimation des paramètres par la **maximisation** de la fonction de vraisemblance
- théorie statistique ancienne et éprouvée, permettant le **test d'hypothèses** par comparaison de modèles
- implique une optimisation sur un **espace mixte** discret / continu
- **PHYML**, un programme performant développé en France (Guindon & Gascuel 2003)

## Approche Bayésienne et MCMC

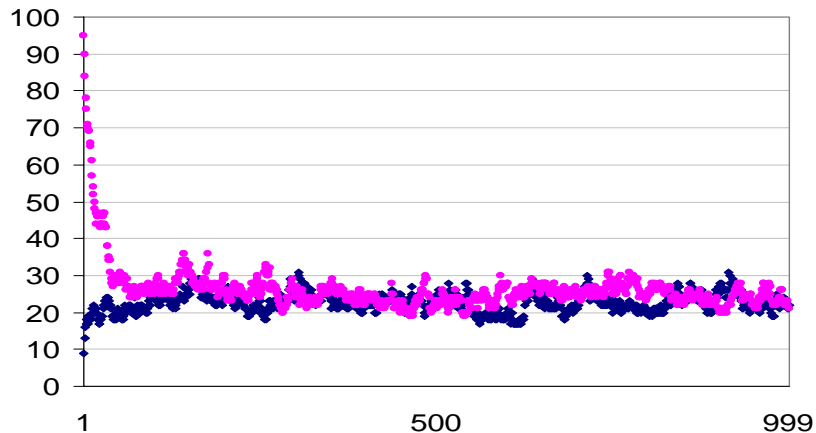
- **théorème de Bayes:**

$$\text{Pr}(\theta \mid \mathbf{D}) = \frac{\text{Pr}(\mathbf{D} \mid \theta) \text{Pr}(\theta)}{\text{Pr}(\mathbf{D})}$$

Diagram illustrating the components of Bayes' theorem:

- vraisemblance** (likelihood) points to  $\text{Pr}(\mathbf{D} \mid \theta)$
- prior** points to  $\text{Pr}(\theta)$
- posterior** points to  $\text{Pr}(\theta \mid \mathbf{D})$
- constante** (constant) points to  $\text{Pr}(\mathbf{D})$

- **échantillonnage** de la distribution postérieure par chaînes de Markov Monte-Carlo



## Approche Bayésienne et MCMC

- **théorème de Bayes:**

$$\Pr(\theta | \mathbf{D}) = \frac{\Pr(\mathbf{D} | \theta) \Pr(\theta)}{\Pr(\mathbf{D})}$$

Diagram illustrating the components of Bayes' theorem:

- vraisemblance** (likelihood) points to  $\Pr(\mathbf{D} | \theta)$
- prior** points to  $\Pr(\theta)$
- posterior** points to  $\Pr(\theta | \mathbf{D})$
- constante** (constant) points to  $\Pr(\mathbf{D})$

- **échantillonnage** de la distribution postérieure par chaînes de Markov Monte-Carlo
- permet l'usage de modèles très **complexes**
- nécessite l'usage d'**a priori** sur la distribution des paramètres

## **Le projet MODEL\_PHYLO: de meilleurs modèles pour de meilleures inférences**

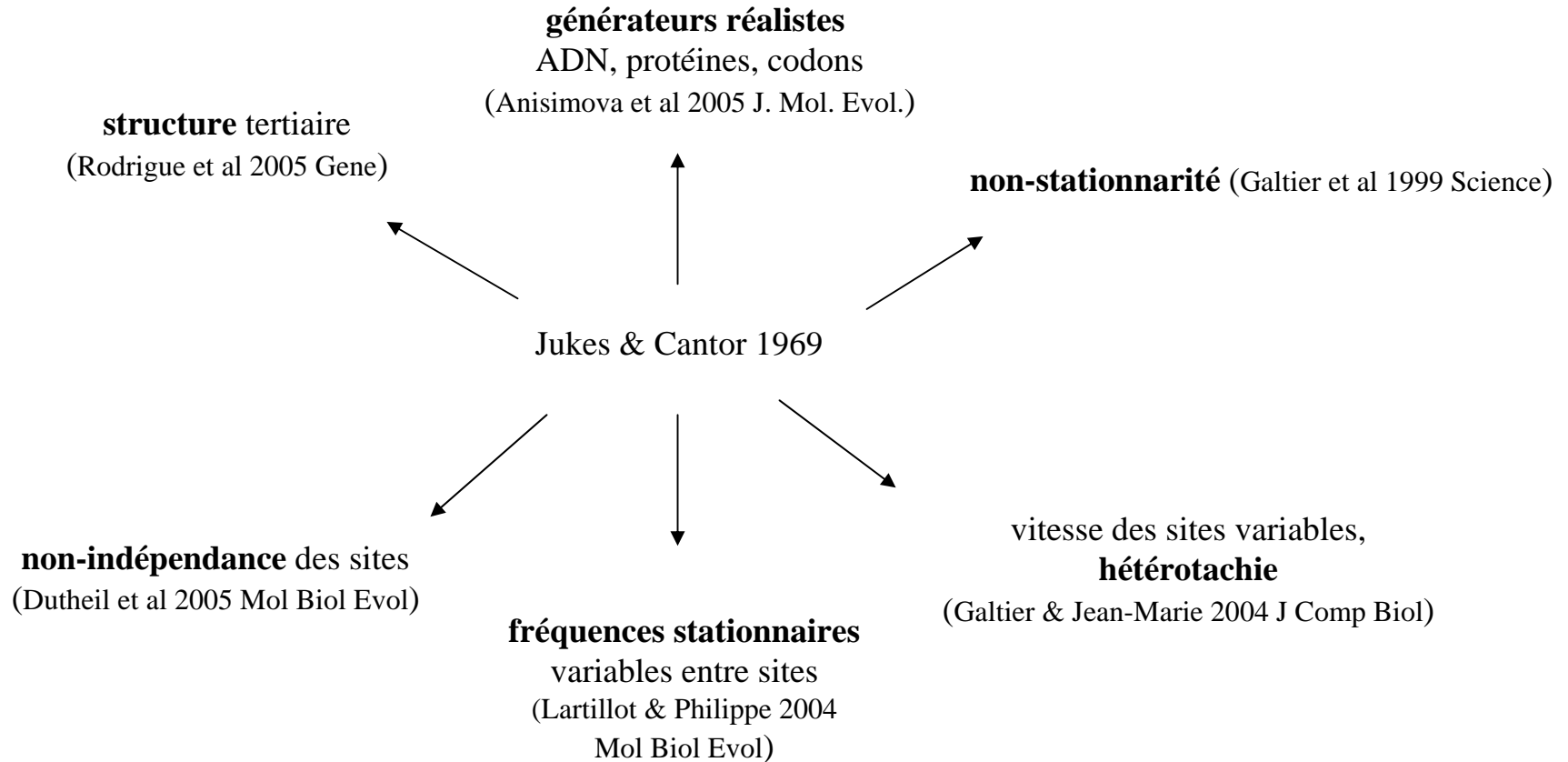
**Objectif:** prendre en compte les particularités des processus de l'évolution moléculaire pour reconstruire l'histoire des gènes et des génomes.

**Approches:** modélisation Markovienne, estimation de paramètres, maximum de vraisemblance, approche bayésienne, développements algorithmiques, bases de données

### **Applications biologiques prévues:**

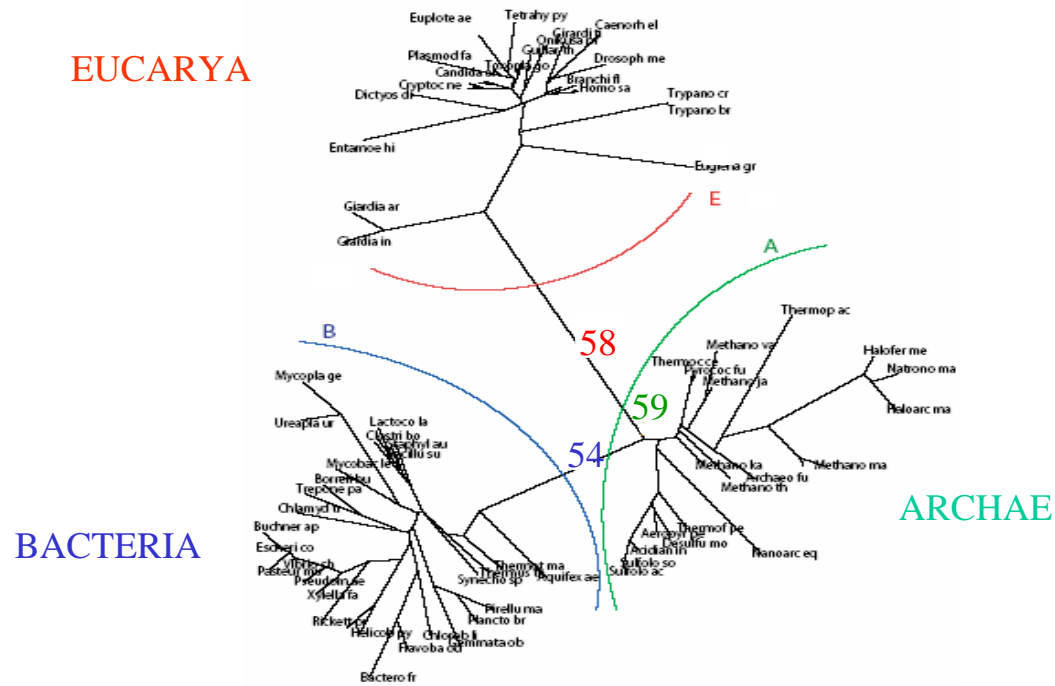
- relations phylogénétiques entre grandes lignées d'eucaryotes et de procaryotes
- évolution de la thermophilie et origine de la vie
- évolution des protéines, contraintes et adaptations

## Modèles Markoviens pour la phylogénie



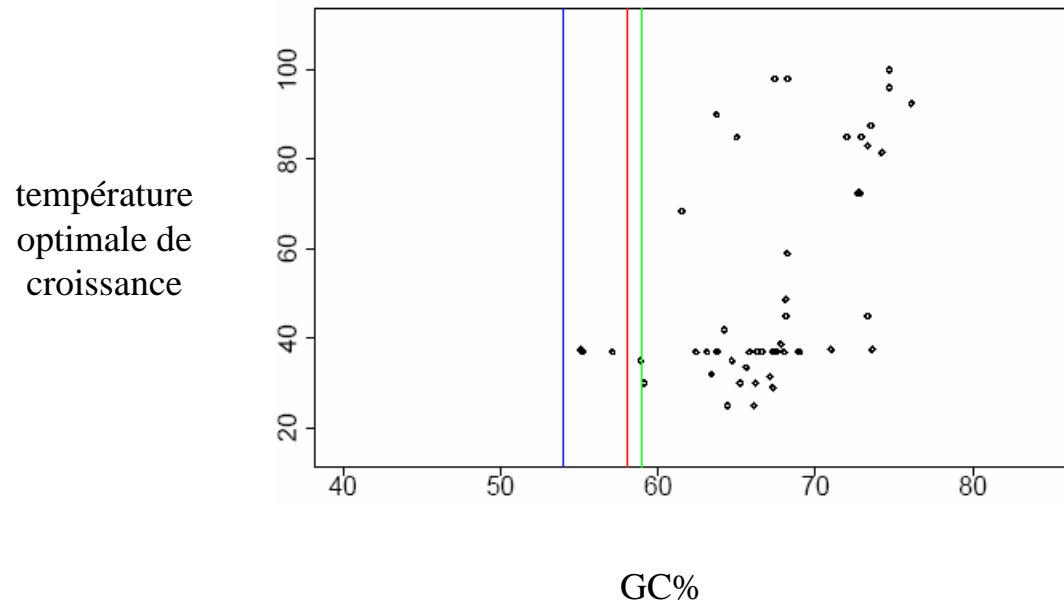
## GC% ancestral de l'ARNr et thermophilie

**NHPHYML**, synthèse de NHML (Galtier & Gouy 1998) et PHYML (Guindon & Gascuel 2003):  
un algorithme rapide pour la reconstruction phylogénétique sous un modèle non-stationnaire.

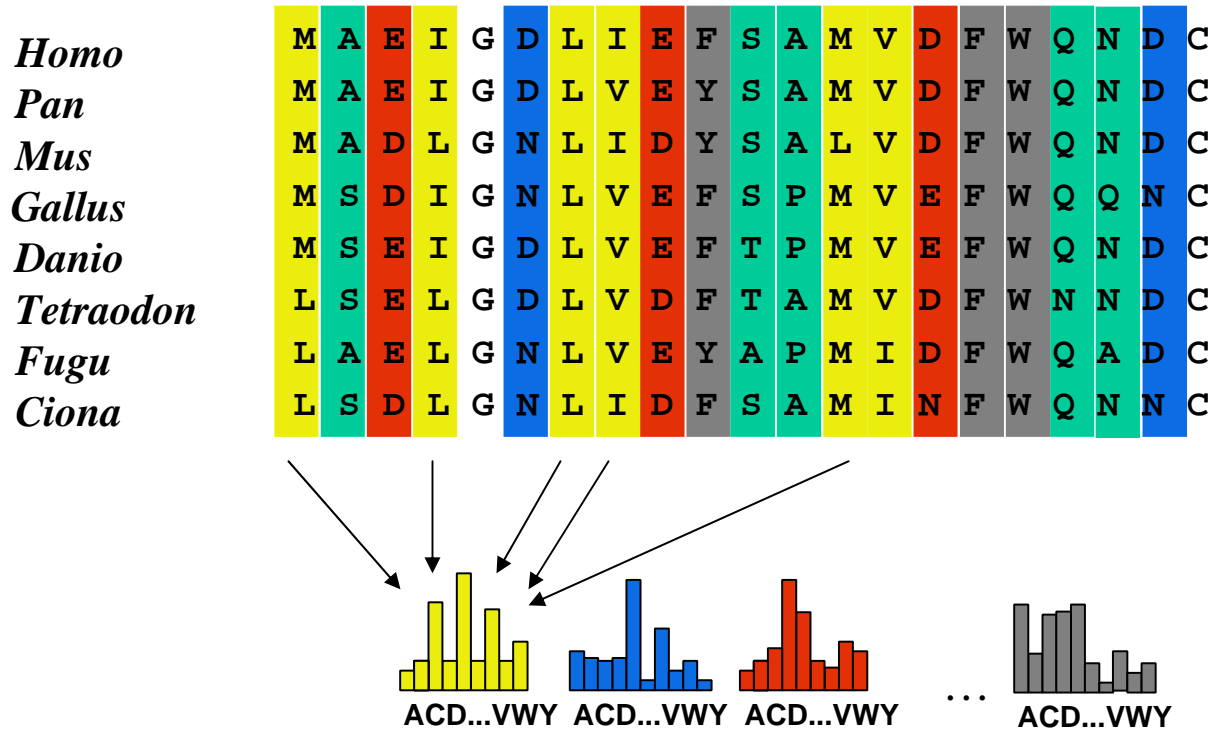


## GC% ancestral de l'ARNr et thermophilie

**NHPHYML**, synthèse de NHML (Galtier & Gouy 1998) et PHYML (Guindon & Gascuel 2003):  
un algorithme rapide pour la reconstruction phylogénétique sous un modèle non-stationnaire

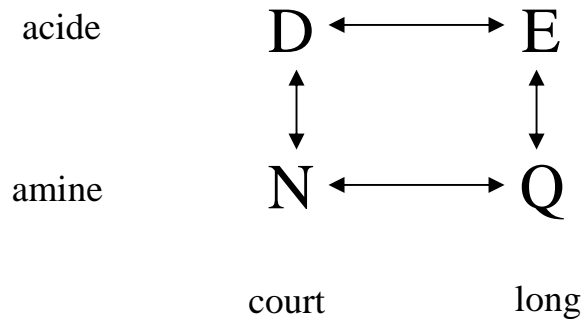


## Variations des fréquences stationnaires entre sites: le modèle CAT



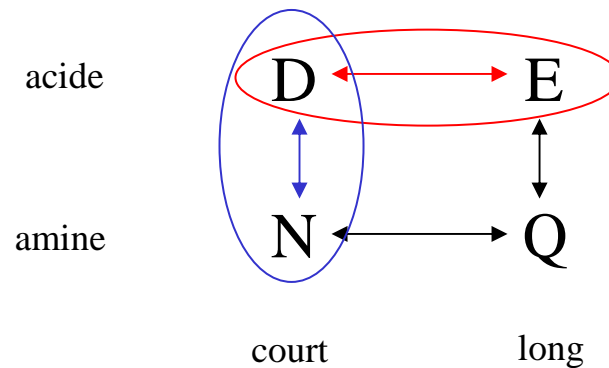
## Variations des fréquences stationnaires entre sites: le modèle CAT

<i>Homo</i>	M	A	E	I	G	D	L	I	E	F	S	A	M	V	D	F	W	Q	N	D	C
<i>Pan</i>	M	A	E	I	G	D	L	V	E	Y	S	A	M	V	D	F	W	Q	N	D	C
<i>Mus</i>	M	A	D	L	G	N	L	I	D	Y	S	A	L	V	D	F	W	Q	N	D	C
<i>Gallus</i>	M	S	D	I	G	N	L	V	E	F	S	P	M	V	E	F	W	Q	Q	N	C
<i>Danio</i>	M	S	E	I	G	D	L	V	E	F	T	P	M	V	E	F	W	Q	N	D	C
<i>Tetraodon</i>	L	S	E	L	G	D	L	V	D	F	T	A	M	V	D	F	W	N	N	D	C
<i>Fugu</i>	L	A	E	L	G	N	L	V	E	Y	A	P	M	I	D	F	W	Q	A	D	C
<i>Ciona</i>	L	S	D	L	G	N	L	I	D	F	S	A	M	I	N	F	W	Q	N	N	C

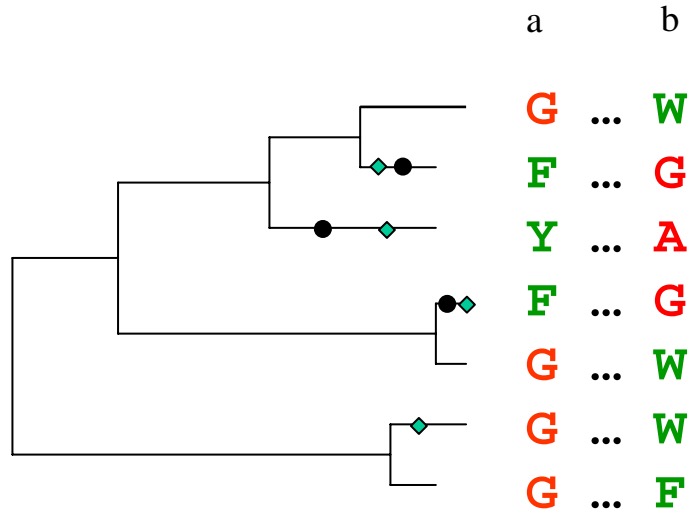


## Variations des fréquences stationnaires entre sites: le modèle CAT

<i>Homo</i>	M	A	E	I	G	D	L	I	E	F	S	A	M	V	D	F	W	Q	N	D	C
<i>Pan</i>	M	A	E	I	G	D	L	V	E	Y	S	A	M	V	D	F	W	Q	N	D	C
<i>Mus</i>	M	A	D	L	G	N	L	I	D	Y	S	A	L	V	D	F	W	Q	N	D	C
<i>Gallus</i>	M	S	D	I	G	N	L	V	E	F	S	P	M	V	E	F	W	Q	Q	N	C
<i>Danio</i>	M	S	E	I	G	D	L	V	E	F	T	P	M	V	E	F	W	Q	N	D	C
<i>Tetraodon</i>	L	S	E	L	G	D	L	V	D	F	T	A	M	V	D	F	W	N	N	D	C
<i>Fugu</i>	L	A	E	L	G	N	L	V	E	Y	A	P	M	I	D	F	W	Q	A	D	C
<i>Ciona</i>	L	S	D	L	G	N	L	I	D	F	S	A	M	I	N	F	W	Q	N	N	C

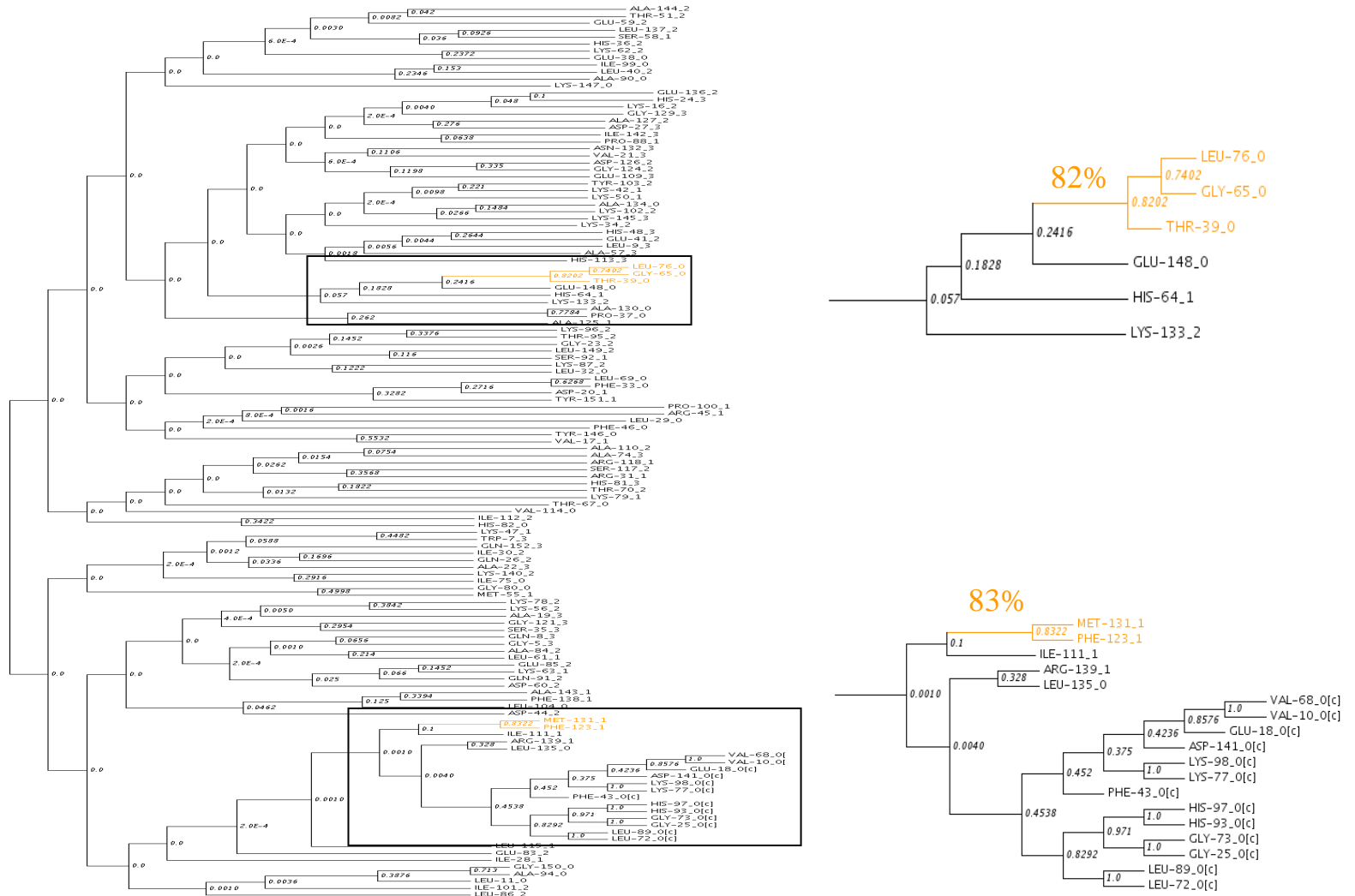


## Cartographie des substitutions et coévolution

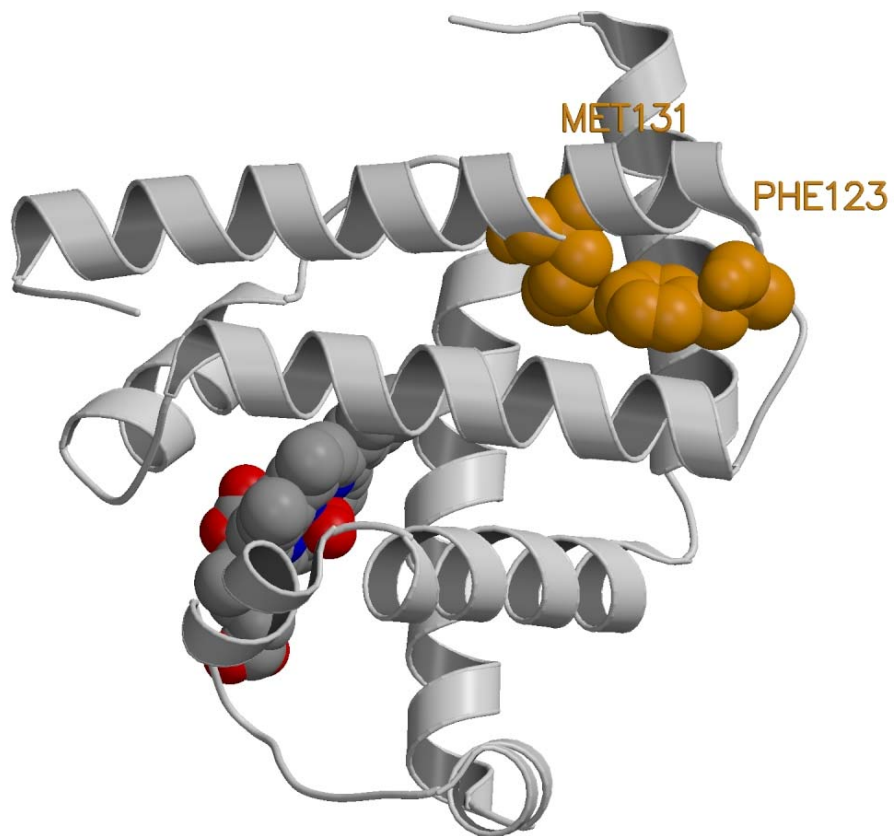


- **cartographie probabiliste** des événements de substitution site-spécifiques
- **classification automatique** des sites sur la base de leur carte de substitution
- évaluation de la robustesse des groupes formés par **bootstrap sur les branches**

# Cartographie des substitutions et coévolution



## Cartographie des substitutions et coévolution



## Cartographie des substitutions et coévolution

