

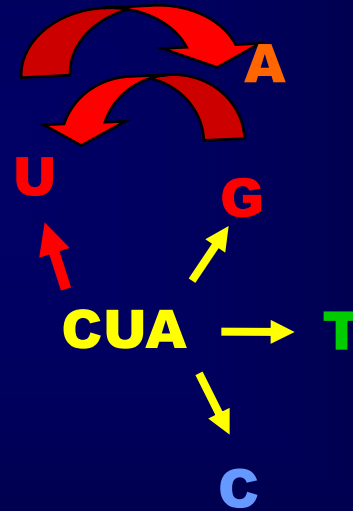
Modèles d'évolution des codons

La plupart des modèles d'évolution des nucléotides assument que les nucléotides évoluent de manière indépendante

Pour les séquences codantes, ce n'est pas le cas!

Leu **CUA**
 CUG
 CUT
 CUC

 UUA
 UUG



Le nombre de substitutions synonymes est limité et les possibilités dépendent l'une de l'autre.

Pour tenir compte de la liaison entre les 3 sites d'un codon, des modèles d'évolution qui définissent le codon comme unité indépendante ont été développées.

Ces modèles sont complexes et riches en paramètres.

Ils ne sont donc jamais utilisés pour effectuer des recherches d'arbres.

A quoi ils servent alors...!?

Les « codon-based models » servent à évaluer précisément

- les taux de substitutions
- les taux de transitions vs transversions
- la pression de sélection
- ...

sur des séquences codantes.

A Codon-based Model of Nucleotide Substitution for Protein-coding DNA Sequences

*Nick Goldman** and *Ziheng Yang†*

*Laboratory of Mathematical Biology, National Institute for Medical Research; and †Biometrics Section, Department of Zoology, The Natural History Museum

A codon-based model for the evolution of protein-coding DNA sequences is presented for use in phylogenetic estimation. A Markov process is used to describe substitutions between codons. Transition/transversion rate bias and codon usage bias are allowed in the model, and selective restraints at the protein level are accommodated using physicochemical distances between the amino acids coded for by the codons. Analyses of two data sets suggest that the new codon-based model can provide a better fit to data than can nucleotide-based models and can produce more reliable estimates of certain biologically important measures such as the transition/transversion rate ratio and the synonymous/nonsynonymous substitution rate ratio.

Mol. Biol. Evol. 11(5):725-736. 1994.

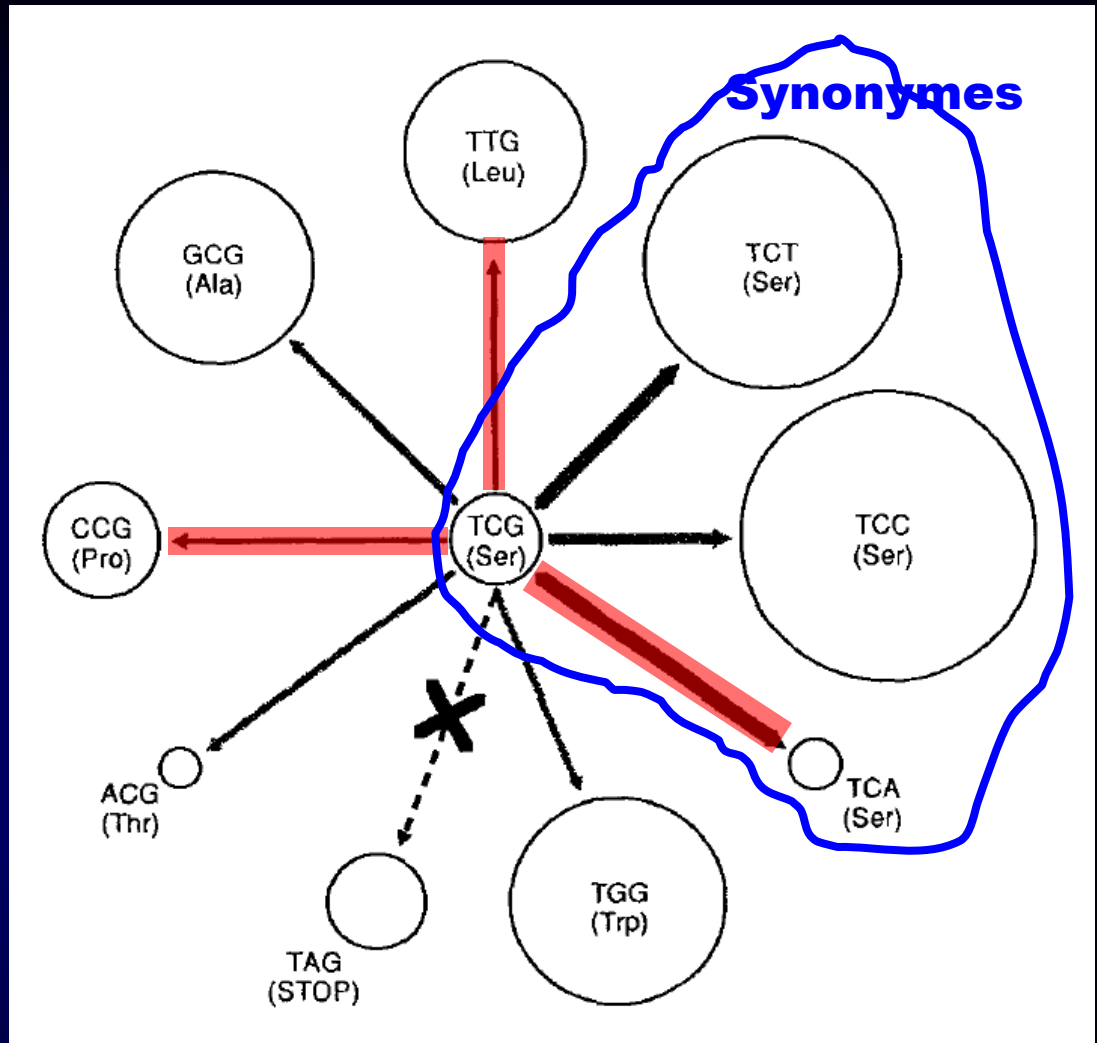
© 1994 by The University of Chicago. All rights reserved.

0737-4038/94/1105-0002\$02.00

Les 9 voisins de TCG

Diamètre = fréquence empirique du codon

Rouge = Transitions
Noir = Transversions



A Codon-based Model of Nucleotide Substitution for Protein-coding DNA Sequences

Nick Goldman* and Ziheng Yang†

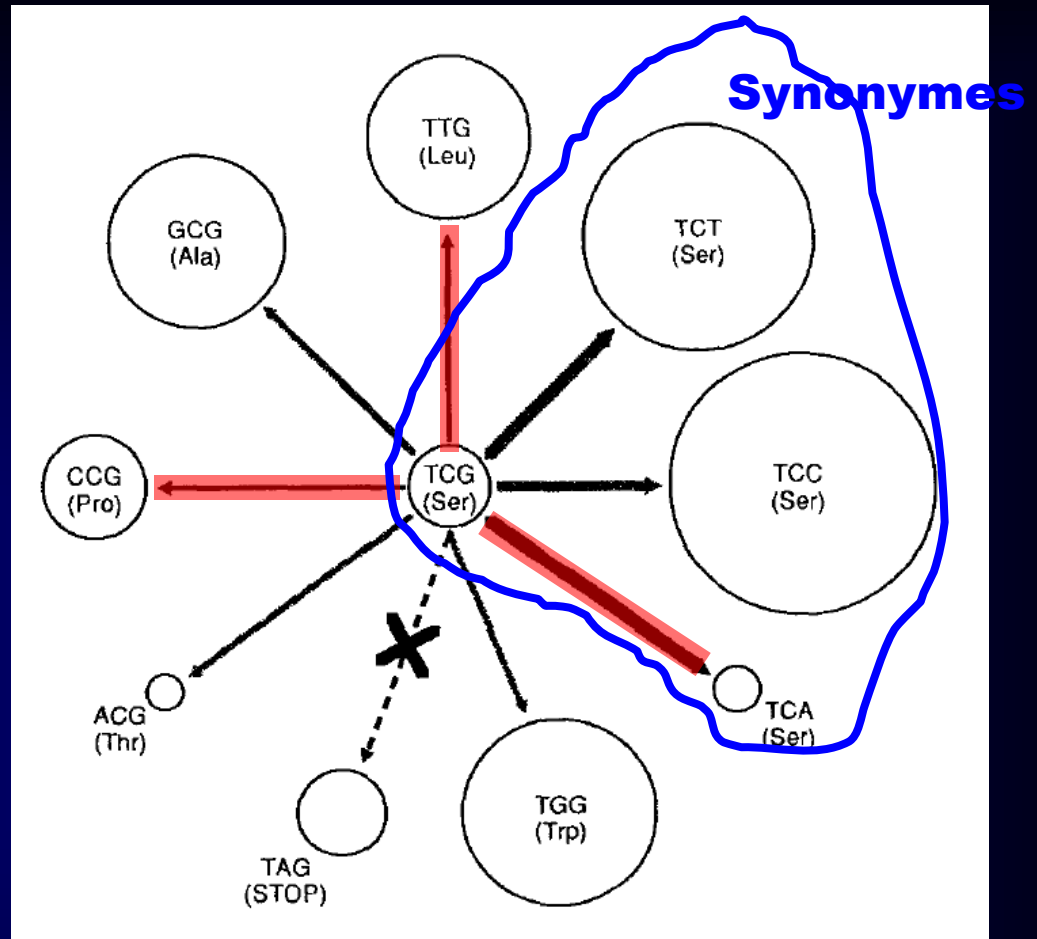
Mol Biol Evol 1994, 11:725-736

FIG. 1.—Example of the “neighbors” to which a codon (here, TCG) may evolve instantaneously through a nucleotide substitution at one position. TCG has eight neighbors, substitution of A for C at the second position being disallowed, as it results in the stop codon TAG. Transitions are marked with black arrows, transversions with gray arrows. Substitutions involving no change in amino acid (generally occurring at a higher rate in this model) are marked with thicker arrows. The size of each circle (except the stop codon TAG) represents the (equilibrium) frequency of that codon, in this case taken from the pooled α - and β -globin gene sequences.

Les 9 voisins de TCG

Diamètre = fréquence empirique du codon

Rouge = Transitions
Noir = Transversions

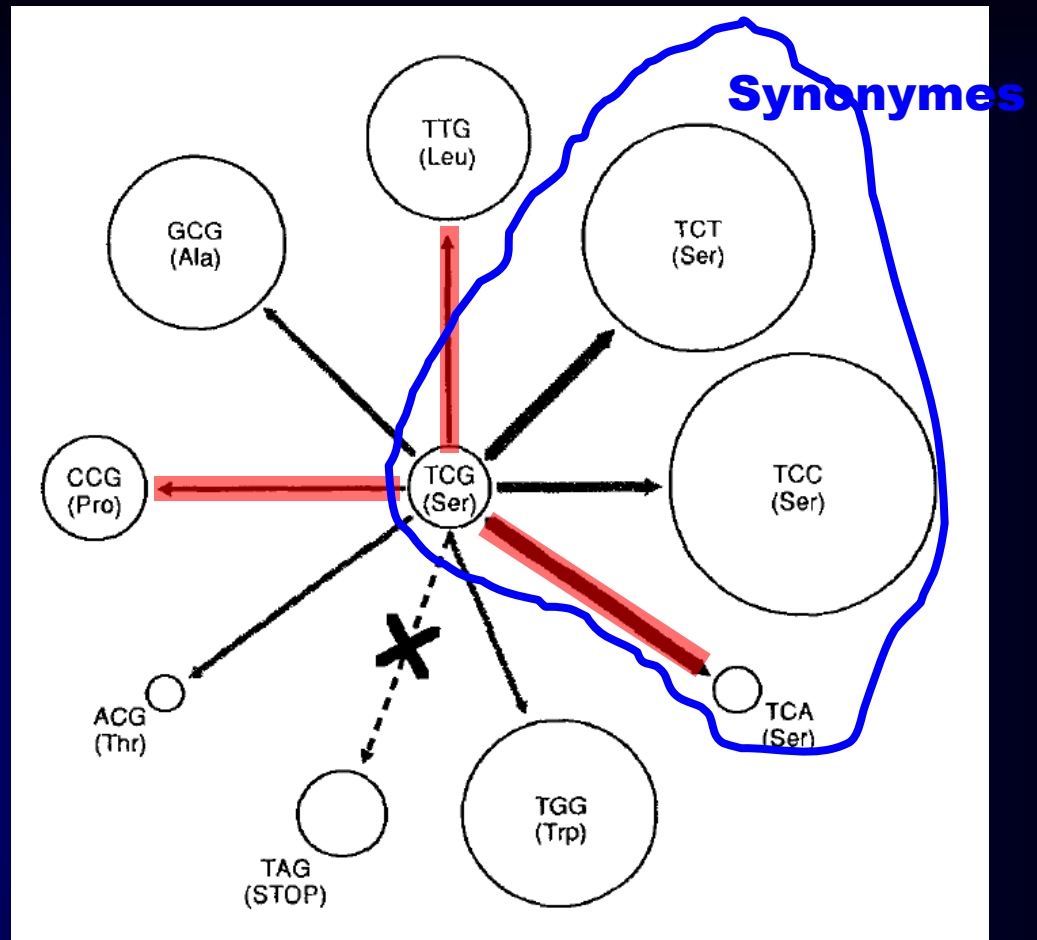


1) Le taux de chaque substitution est proportionnel à la fréquence du codon d'arrivée (diamètre)

Les 9 voisins de TCG

Diamètre = fréquence empirique du codon

Rouge = Transitions
Noir = Transversions

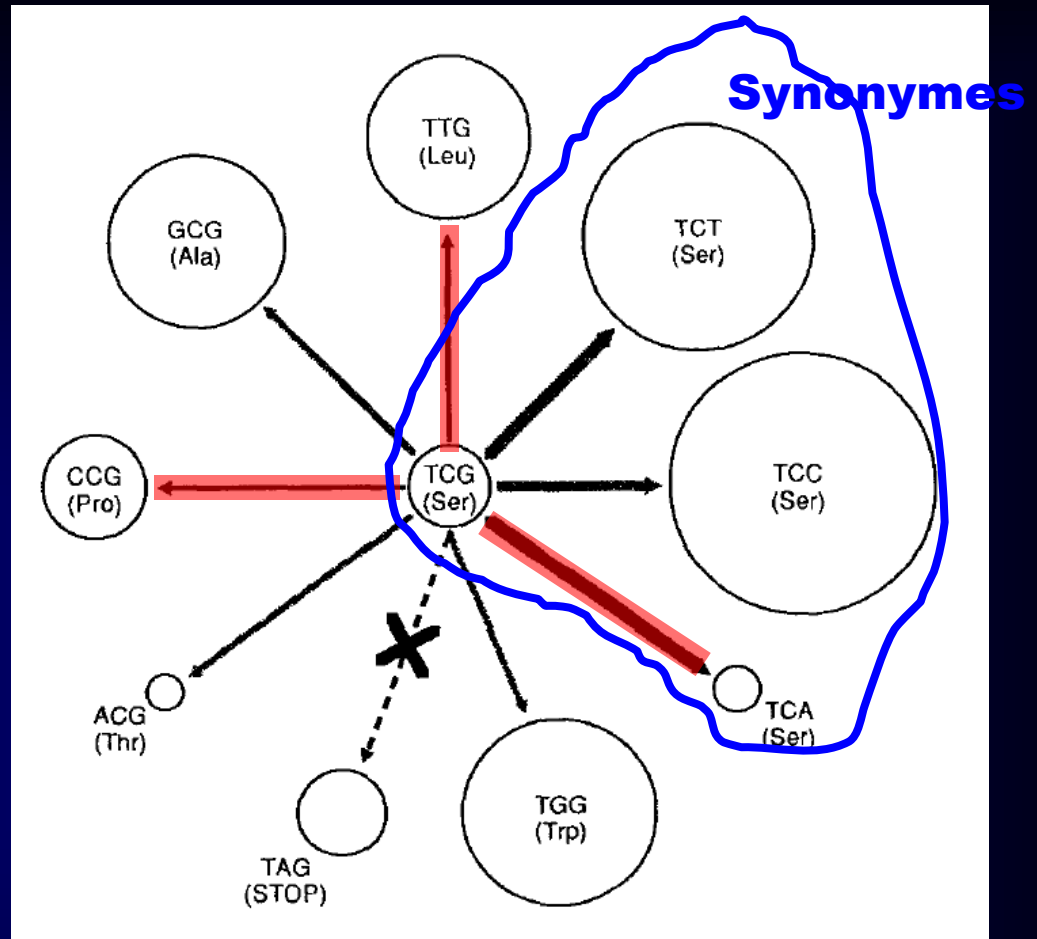


2) Les transitions sont pondérées comme dans le Modèle K2P

Les 9 voisins de TCG

Diamètre = fréquence empirique du codon

Rouge = Transitions
Noir = Transversions



3) Pondère les cas où l'acide aminé est changé en suivant une matrice de substitution des AA (ici: matrice de Grantham, 1974)

Généralisation d'un modèle de base qui ne fait pas appel à une matrice empirique (Yang et al., 1998)

others are very different. This simplified version (Yang et al. 1998) specifies the substitution rate from codon i to codon j as

$$q_{ij} = \begin{cases} 0, & \text{if the two codons differ at more than one position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition,} \end{cases}$$

Fréq. d'équilibre
Pondération par κ
Pondération par ω
Pondération par κ et ω

The equilibrium frequency of codon j (π_j) can be considered a free parameter, but can also be calculated from the nucleotide frequencies at the three codon positions (control variable `CodonFreq`). Under this model, the relationship holds that $\omega = d_N/d_S$, the ratio of nonsynonymous/synonymous substitution rates. This basic model is fitted by specifying model = 0 `NSsites = 0`, in the control file `codeml.ctl`. It forms the basis for more sophisticated models implemented in `codeml`.

κ =transitions/transversions

ω = taux subst.non-synonymes/taux subst. synonymes

Quel programme implémente ce type de modèle ?

Paml (Ziheng Yang 1997)

Phylogenetic Analysis by Maximum Likelihood

Présentation de Paml

Le package Paml n'est pas vraiment destiné à faire des recherches d'arbres car les modèles d'évolution qu'il propose sont complexes et donc très couteux en temps de calcul.

Les nombreux modèles permettent d'effectuer une multitude de tests.

- Estimations de paramètres pour des modèles complexes, avec partitions, et taux d'évolution variable entre sites**
- Likelihood ratio tests, tests de topologies et tests d'hypothèses**
- Estimation des temps de divergence avec horloge moléculaire Globale ou locale**
- Reconstruction des états (séquences) ancestraux**
- Estimation du taux de substitutions synonymes (dS) et non-synonymes (dN) et calcul des pressions de sélection**
- Simulations d'évolutions**

Présentation de Paml

Format reconnu: Phylip (alignements et arbres)

Deux principaux programmes:

- **Baseml (non-codant)**
- **Codeml (codant)**

Pression de sélection

Le taux K_s/K_a est utilisé pour évaluer la pression de sélection sur les régions codantes.

$K_A/K_s \approx 1$ -> les séquences évoluent de manière neutre, sans sélection.

$K_A/K_s > 1$ -> La région codante est sous sélection positive.

$K_A/K_s \ll 1$ -> La région codante est sous sélection purificatrice.

$$K_a = dN$$

$$K_s = dS$$

$$K_a/K_s = dN/dS = \omega$$

Paml

Le modèle de base et ses variantes permettent d'estimer la pressions de sélection dN/dS soit:

- comme valeur moyenne pour tout l'arbre (modèle de base)

- pour une branche en particulier vs le reste des branches

- branche par branche

(branch models)



Paml

Il est aussi possible de calculer la pression de sélection qu'à subit chaque codon de l'alignement.

Sites models

Model	NSsites	np	Free parameters
M0 (one ratio)	NSsites = 0	1	ω
M1a (NearlyNeutral): p_0 ($p_1 = 1 - p_0$) $\omega_0 < 1, \omega_1 = 1$	NSsites = 1	2	$p_0, \omega_0 < 1$
M2a (PositiveSelection): p_0, p_1 ($p_2 = 1 - p_0 - p_1$) $\omega_0 < 1, \omega_1 = 1, \omega_2 > 1$	NSsites = 2	4	$p_0, p_1,$ $\omega_0 < 1, \omega_2 > 1$
M3 (discrete): p_0, p_1 ($p_2 = 1 - p_0 - p_1$) $\omega_0, \omega_1, \omega_2$	NSsites = 3	5	$p_0, p_1,$ $\omega_0, \omega_1, \omega_2$
M7 (beta): p, q	NSsites = 7	2	p, q
M8 (beta& ω): p_0 ($p_1 = 1 - p_0$) $p, q, \omega_s > 1$	NSsites = 8	4	$p_0, p, q, \omega_s > 1$

LRT

LRT

Distribution Beta varie de 0 à 1 et est déterminée par p et q

Exercices