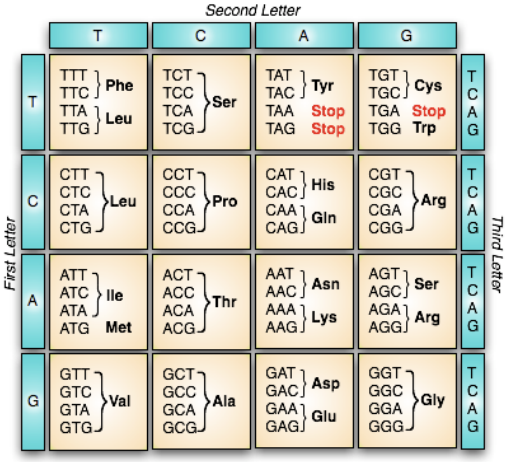
TP 1 – Évolution moléculaire

Nicolas Lartillot
nicolas.lartillot@univ-lyon1.fr

15 novembre 2017

1 Taux de transition/transversion et dN/dS

L'objectif de ce premier exercice est d'effectuer un calcul empirique simple (et quelque peu approximatif) du taux de transition sur transversion et du dN/dS chez deux couples d'espèces (deux primates, deux rongeurs). Ce calcul vise à illustrer un problème particulièrement pernicieux dans l'estimation empirique de ces deux quantités (et plus généralement de l'estimation des taux de substitution/mutation sur les séquences codantes).



On possède deux alignements de séquences codantes, le premier entre l'homme et le macaque, le second entre le rat et la souris. On fera dans un premier temps les calculs sur le rat et la souris.

Question 1 *Calculer le nombre total de différences nucléotidiques entre les deux espèces, ainsi que la fraction des positions qui sont différentes.*

Question 2 *Séparez dans vos décomptes les transitions (entre C et T ou entre G et A) et les transversions (toutes les autres).*

On suppose que le processus de mutation est tel que chaque mutation de type transversion a lieu à un taux u par génération, tandis que chaque mutation de type transition se produit à un taux κu par génération. Par exemple, si $\kappa = 2$, alors les mutations de C vers T sont deux fois plus fréquentes que les mutations de C vers A.

Question 3 *En chaque position, combien de mutations sont possibles ? Combien d'entre elles sont des transitions ? Combien d'entre elles sont des transversions ? Donnez une estimée de κ.*

On considère une petite partie de l'alignement, ci-dessous :

```
... CAC TTC CGC TTC GTG ...
... CAT TTC CGT TCC GTG ...
```

Question 4 *Comptez le nombre de différences synonymes et non-synonymes entre les deux séquences*

Question 5 *Parmi les 45 plus proches voisins (qui diffèrent par un seul changement nucléotidique) de la séquence de la souris, combien sont synonymes, et combien sont non-synonymes ? Quelle serait votre estimée du dN/dS ?*

Pour automatiser ce calcul sur de plus grandes séquences, vous disposez d'un script R, `gencode.R`, qui code pour diverses fonctions prédéfinies permettant des manipulations prenant en compte le code génétique. En particulier, on a les deux fonctions suivantes :

- `neighborsSynNonSyn(s)` : cette fonction calcule le nombre de voisins synonymes et non-synonymes pour un codon donné `s`. Par exemple, `neighborsSynNonSyn('AAA')` renvoie la paire $(1, 7)$ (1 voisin synonyme et 7 voisins non-synonymes pour le codon AAA)
- `diffSynNonSyn(s1,s2)` : pour deux codons qui ne diffèrent que d'un nucléotide, cette fonction renvoie une paire, qui sera égale à $(1, 0)$ si les deux codons sont synonymes et $(0, 1)$ si les deux codons sont non-synonymes. Par exemple, `diffSynNonSyn('AAA', 'AAT')` renvoie $(0, 1)$, puisque AAA code pour la lysine et AAT pour l'asparagine.

Question 6 *Comptez les nombre de différences synonymes et non-synonymes entre les deux séquences. Comptez également le nombre total de mutations possibles, synonymes et non-synonymes, en partant de la séquence de la souris. Donnez une première estimée du dN/dS.*

Reprenons maintenant l'exemple de l'alignement ci-dessus (de 5 codons).

Question 7 *Comptez séparément, à la main, le nombre de différences synonymes et non-synonymes en transition et en transversion entre les deux séquences. Comptez également le nombre total de mutations possibles, synonymes et non-synonymes, en transition ou en transversion, en partant de la séquence de la souris.*

Pour automatiser ce calcul sur de plus grandes séquences, vous disposez des fonctions suivantes :

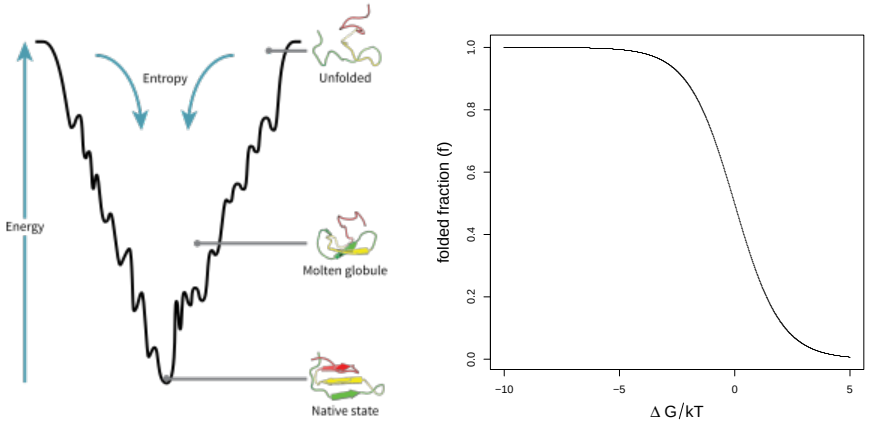
- `neighborsSynNonSynTsTv(s)` : cette fonction renvoie un quadruplet, donnant dans l'ordre le nombre de voisins synonymes en transition, en transversion, puis le nombre de voisins non-synonymes en transition, puis enfin en transversion. Par exemple, `neighborsSynNonSynTs('AAA')` renvoie $(1, 0, 2, 5)$.
- `diffSynNonSynTsTv(s1,s2)` : pour deux codons qui ne diffèrent que d'un nucléotide, cette fonction renvoie un quadruplet, qui sera égal à $(1, 0, 0, 0)$ si les deux codons sont synonymes en transition, $(0, 1, 0, 0)$ si synonymes en transversion, $(0, 0, 1, 0)$ si non-synonymes en transition et $(0, 0, 0, 1)$ si non-synonymes en transversion.

Question 8 *Recalculez κ sur les changements synonymes uniquement. Comparer cette estimée à celle effectuée ci-dessus. Laquelle est la plus fiable, en tant que mesure du taux de transition/transversion mutationnel ?*

Question 9 *Calculez séparément le dN/dS sur les transitions et les transversions. Que constatez-vous par rapport au dN/dS estimée plus haut ? Comment expliquez vous cette différence ?*

Question 10 *Combinez vos deux estimées de dN/dS en transition et en transversion en une seule estimée finale. Pour cela, vous utiliserez votre estimée de κ. Effectuez le même calcul chez l'humain et le macaque.*

2 Stabilité conformationnelle des protéines



Ce deuxième exercice vise à construire un modèle de simulation très simplifié de l'évolution d'une protéine sous la contrainte sélective liée au maintien de sa stabilité conformationnelle. On cherchera plus précisément à quantifier le nombre moyen de mutations délétères, ainsi que la fraction mal repliée de la protéine, à l'équilibre mutation-sélection-dérive, pour différentes tailles de population et différents niveaux d'expression.

On considère une protéine qui possède une configuration native bien définie. L'énergie libre de repliement de cette protéine est noté ΔG. D'après la thermodynamique, à la temperature ambiante T, la fraction de la protéine qui est incorrectement repliée est égale à :

$$f = \frac{1}{1 + e^{\frac{-\Delta G}{kT}}}$$

Typiquement, les ΔG des protéines naturelles sont négatifs, par conséquent, la fraction mal repliée est en générale assez faible.

Dans notre cas, on imagine qu'il existe une séquence optimale pour la protéine, pour laquelle $\Delta G_0 = -10kT$. Toutefois, des mutations délétères peuvent intervenir. Ces mutations ont tendance à rendre la protéine moins stable à l'équilibre. Pour modéliser ceci, on suppose que la protéine possède $L = 200$ positions en lesquelles les mutations sont déstabilisantes. Chacune de ces mutations ajoute une valeur positive $\Delta\Delta G = 0.1kT$ au ΔG de la protéine. De ce fait, une version de la protéine qui possède un total de m mutations délétères aura un $\Delta G = \Delta G_0 + m\Delta\Delta G$, et la fraction mal repliée pour cette protéine mutante sera alors égale à :

$$f(m) = \frac{1}{1 + e^{b - am}}$$

avec $b = -\Delta G_0/kT = 10$ et $a = \Delta\Delta G/kT = 0.1$. Cette fonction f est croissante avec m.

Question 11 *Calculez la valeur de m pour laquelle la fraction mal repliée est égale à 0.5*

Afin de s'affranchir des complications liées au code génétique, on travaillera directement sur la séquence en acides aminés. On imagine par ailleurs qu'il existe une seule séquence optimale pour cette protéine. En chacune des L positions potentiellement déstabilisantes, on a en moyenne 6 voisins non-synonymes, qui sont tous délétères. On suppose donc, pour simplifier, que, en une position donnée, si l'acide aminé actuel est optimal, alors on mute vers un acide aminé non optimal à un taux $6u$, et inversement, si l'acide aminé actuel n'est pas optimal, alors on mute vers l'acide aminé optimal avec un taux u.

Enfin, on fait l'hypothèse que la fitness de l'organisme est une fonction décroissante de la fraction mal repliée :

$$W(m) \quad = \quad e^{-\alpha f}$$

On prendra $\alpha = 1$.

Question 12 *Calculez u_m, le taux de mutation total avec lequel une séquence ayant m mutations délétères mute vers une séquence ayant $m+1$ mutations délétères*

Question 13 *Pour $m > 0$, calculez v_m, le taux de mutation total avec lequel une séquence ayant m mutations délétères mute vers une séquence ayant $m-1$ mutations délétères*

Question 14 *Calculez le coefficient de sélection $s_u(m)$ associé à une mutation délétère survenant sur une protéine ayant déjà m mutations délétères.*

Question 15 *Pour $m > 0$, calculez le coefficient de sélection $s_v(m)$ associé à une mutation faisant passer le nombre de mutations délétères de m à $m-1$.*

Pour calculer ces coefficients de sélection, plutôt que la formule vue en cours, on utilisera une formule alternative (et très proche lorsque s est petit) : $s = \ln W' - \ln W$, où W' est la fitness du mutant et W celle du sauvage.

Question 16 *Si, à un instant t la protéine possède m mutations délétères, quel est le taux de substitution $r_u(m)$ vers une séquence avec $m+1$ mutations délétères ? Quel est le taux de substitution $r_v(m)$ vers une séquence avec $m-1$ mutations délétères ?*

Nous sommes maintenant prêt pour effectuer une simulation. On mesure le temps en millions d'années, on prend $u = 1$ (1 mutation par position et par millions d'années). On prendra comme taille efficace $N = 100000$ dans un premier temps. On part d'une protéine avec $m = 1$, et au temps $t = 0$. Puis, sachant la valeur courante de m :

- on calcule les taux de substitution $r_u(m)$ et $r_v(m)$, ainsi que leur somme $r = r_u(m) + r_v(m)$
- on tire un temps d'attente distribué exponentiellement, de taux r : $\delta t \sim Exp(r)$. Le prochain évènement de substitution a lieu au temps $t + dt$
- cet évènement sera une substitution vers $m+1$ avec une probabilité $r_u(m)/r$, et sinon, sera une substitution vers $m-1$
- on met à jour la valeur de m ; on pose $t = +\delta t$;
- on répète cette procédure jusqu'à avoir simulé un temps total d'un milliard d'années

Question 17 *Programmer le simulateur en R. Visualisez la trajectoire évolutive de la protéine. En particulier, visualisez, au cours du temps, l'évolution du nombre de mutations délétères présentes dans la protéine. Combien de temps faut-il environ pour atteindre l'équilibre mutation-sélection ?*

Question 18 *Calculez le nombre moyen de mutations délétères à l'équilibre, ainsi que la fraction moyenne mal repliée. Faites ce calcul pour* $N = 100000$ *et pour* $N = 10000$*. Que constatez vous ?*

Question 19 *Que deviendrait d'après vous le nombre moyen de mutations délétères chez un organisme hyperthermophile (toutes choses égales par ailleurs). Note : augmenter la température revient à diminuer les valeurs des paramètres* a *et* b*.*

On cherche enfin à obtenir une estimée rapide, approchée, de la valeur moyenne de m à l'équilbre mutation-sélection-dérive.

Question 20 *Calculez, en fonction de* m*, le rapport des taux de mutation de* m *vers* $m+1$ *et de* $m+1$ *vers* m*. On notera ce rapport* λ_m*.*

Question 21 *Calculez, en fonction de* m*, le rapport des probabilités de fixation de* $m+1$ *vers* m *et de* m *vers* $m+1$*. On notera ce rapport* μ_m

Question 22 *Sur un graphe, dessinez à la fois* $\ln \lambda_m$ *et* $\ln \mu_m$ *en fonction de* m*, pour* m *variant de 1 à 50.*

Question 23 *En déduire la valeur moyenne d'équilibre de* m

Références

Taverna, D. M., and Goldstein, R. A. (2002). Why are proteins marginally stable ? Proteins, 46(1), 105 ?109.

Goldstein, R. A. (2011). The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins, 79(5), 1396 ?1407. http ://doi.org/10.1002/prot.22964