M1 Bionformatique / Biotechnologie

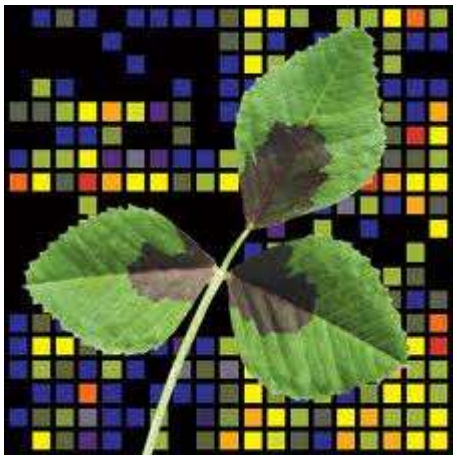
TP Evolution Moléculaire : Détection moléculaire de l'adaptation

Bien que le concept Darwinien d'adaptation ait été formalisé il y a un siècle par les généticiens (Fisher, Wright, Haldane), il a été difficile de démontrer rigoureusement que des différences en acide aminés entre protéines homologues avaient une signification adaptative. Depuis plus de 30 ans, les généticiens des populations et les théoriciens de l'évolution développement des méthodes pour détecter la signature de la sélection naturelle à l'échelle des séquences d'ADN (codant mais aussi non codant). Grâce à l'avènement des technologies de séquençage à haut-débit, un grand nombre de données de séquences peuvent maintenant être explorées pour évaluer l'influence de la sélection naturelle dans l'évolution des populations et des espèces.

Ce TP vise à explorer 3 types d'approches permettant de détecter la signature de la sélection naturelle. Ces méthodes dépendent du type d'échantillonnage des séquences d'ADN (alignement intra/interspécifiques) et du type de séquence (codante, non codantes). Nous utiliserons 2 logiciels : DNAsp (Librado and Rozas, 2009) et Mega (Kumar and Tamura et al., 2011).

Partie 1

L'espèce ***Medicago truncatula*** est une plante légumineuse modèle très étudiées en laboratoire, notamment dans le domaine des interactions plantes-microorganismes. En effet, cette espèce est capable d'établir une symbiose avec des bactéries fixatrices d'azote (*Rhizobium*) et des champignons endomycorhiziens à arbuscles (*Glomus intraradices*). Cette espèce est aussi la cible de microorganismes pathogènes. (ex : *Aphanomyces euteiches*). ***Medicago truncatula*** **est donc un bon modèle pour rechercher des gènes impliqués dans l'adaptation de la plante à différents microorganismes**. Différentes accessions (lignées) de *Medicago truncatula*_ ont été séquencées par des méthodes haut-débit (séquençage 454). L'alignement des lectures (reads) sur le génome de référence (lignée A17) à permis de détecter des polymorphismes de séquences à l'échelle d'une base (Single Nucleotide Polymorphisms= SNPs).



Le travail consiste à effectuer un « genome scan » pour repérer les régions du génome soumises à des pressions de sélection. Nous travaillons ici sur **une petite région du génome séquencée chez plusieurs accessions (individus) de l'espèce *Medicago truncatula***.

- Ouvrir le fichier « Medicago_SNPs_sequences.fasta » dans DNAsp
  - combien de sites polymorphes étudions-nous ?
    256 (longueur totale 803, dont 547 monomorphes)
- Effectuez une analyse du polymorphisme ADN (globale et par fenêtres glissantes. Utiliser plusieurs paramètres de fenêtres glissantes) :
  - Que signifient **S**, **Theta (per sequence) from S**, **k**, **Theta (per site) from S**, et **Nucleotide diversity π (Pi)**, et sur quoi vous renseignent-ils ?
    Valeurs sur la séquence entière :
    S = nombre de sites polymorphes (nombre de sites mutés) = 256
    θ_S = **Theta (per sequence) from S**, Theta-W = 49,446
    θ_π = Average number of nucleotide differences, **k**: 59,738

$$\theta_s = S \Big/ \sum_{i=1}^{n-1} 1/i \qquad\qquad \theta_\pi = \binom{n}{2}^{-1} \sum_{i<j}^{n-1} d_{ij}$$

    Le dénominateur de θ_S est une correction pour la taille de l'échantillon (i.e. nombre de séquences). Pour θ_π, d_{ij} est le nombre de différences nucléotidiques entre 2 séquences, calculée pour toutes les séquences et on fait la moyenne sur toutes les paires de séquences ($\binom{n}{2}$).
    Valeurs par nucléotide (« per site »):

$$\frac{\theta_s}{L} \qquad\qquad \frac{\theta_\pi}{L}$$

    Où *L* est la longueur de la séquence
    θ_s = **Theta (per site) from S**, Theta-W: 0,06158 (estimateur de Watterson - 1975)
    θ_π = **Nucleotide diversity, π = Pi**: 0,07439 (diversité nucléotidique : probabilité d'hétérozygotie à un nucleotide)
  - Existe-t-il une variation de ces paramètres le long de cette région du génome ?
    Sliding window (100,25) : π varie (autour de 0.1 de 200 à 600, de 0.04 à 0.08 de part et d'autre). S et θ ont la même courbe (θ est ramené à la longueur de la séquence)
- Avec quel test mettriez-vous en évidence un potentiel effet de la sélection ?
  Test de Tajima : $D = \theta_\pi - \theta_S / SE(\theta_\pi + \theta_S)$. H0=Loi Normale. Explication de H1
- Effectuez le test (global, et par fenêtres glissantes), et interprétez.
  Test de Tajima global (sur les 803): D=0.70096 p-value>0.10
  Sliding window (100,25) : 1.5<D<3 de position 250 à 600 et -1<D<0 de part et d'autre
  Test de Tajima local :
  - de position 1 à 200 D=-0.65 p-value>0.10 (Dans « Tools » simulation coalescent : P[D≤-0.65] = 0.281 = p-value
  - de position 200 à 600 D=2.11 p-value<0.05 (simulation coalescent : P[D≤2.11]=0.98 => p-value=1-0.98=0.02)
  - de position 600 à 803 D=-0.7 p-value>0.10 (simulation coalescent : P[D≤-0.7] = 0.139 = p-value)

D>0 significatif : signature de sélection balancée (allèles maintenus en fréquences intermédiaire = maintien du polymorphisme – cf diapo du cours)

D'après l'annotation du génome de référence, la région génomique que nous étudions est plutôt pauvre en gènes, mais elle comporte un petit cluster de 3 gènes connus pour être impliqués dans la reconnaissance d'effecteurs microbiens (gène de type NBS-LRR).

- Quel lien pouvez-vous faire entre la nature des gènes localisés dans cette région et vos résultats ? Quelle interprétation évolutive pouvez-vous donner ?
  Maintient du polymorphisme au niveau de gènes impliqués dans la reconnaissance d'effecteurs microbiens. Les effecteurs microbiens tentent de contourner les défenses des plantes en limitant, par de nouvelles mutations, la probabilité d'être reconnus par les gènes de reconnaissance de type NBS-LRR. Ces gènes NBS-LRR doivent donc présenter un répertoire allélique large au niveau des sites de reconnaissance des effecteurs de manière à neutraliser le plus possible les effecteurs. La sélection balancée favorise ce maintien, et nous pouvons observer sa signature car les allèles impliqués sont « avantageux » face à tel ou tel effecteur.

Partie 2

La mouche du vinaigre, du genre ***Drosophila***, présente un grand nombre d'espèces qui se sont en partie diversifiées en raison de leur régime alimentaire, ces mouches se nourrissant d'une grande variété de fruits sucrés. Les fruits sucrés contiennent des proportions variables de sucres, et donc d'alcool, au fur et à mesure de leur maturation, jusqu'à leur pourrissement. **L'enzyme Adh (alcool dehydrogenase)** participe à la dégradation de l'alcool qui est toxique pour l'organisme à certaines concentrations. Nous étudions la séquence codante de cette enzyme chez **différents individus de 2 espèces de drosophiles : *Drosophila simulans* et *Drosophila yakuba***. Est-ce que cette enzyme a subi une évolution adaptative chez ces espèces ?

Drosophila simulans* (vigne)** ***Drosophila yakuba

- Ouvrir le fichier « Drosophila_Adh.phy » dans DNAsp. Le format Phylip est un format d'alignement de séquences que l'on peut générer par exemple avec un logiciel d'alignement tel que Clustal X, et peut être lu par différents logiciels d'analyse évolutive des polymorphismes. Sous DNAsp, assignez ces séquences comme une région codante.
- A quels types de mutations s'intéresse-t-on ? où se situent-elles sur un codon donné ?
  Synonymes (essentiellement 3ème base du codon), non synonymes (1ere base du codon)
- Quel test basé sur la comparaison du polymorphisme et de la divergence entre espèces pourriez-vous utiliser afin d'identifier une éventuelle évolution adaptative de l'Adh ?
  Test de Mc Donald et Kreitman sur la comparaison du polymorphisme et de la divergence entre espèces, pour les substitutions synonymes et non synonymes (DNAsp). Test de fisher sur table de contingence :

	Fixées	Polymorphes
S	17	22
NS	6	0

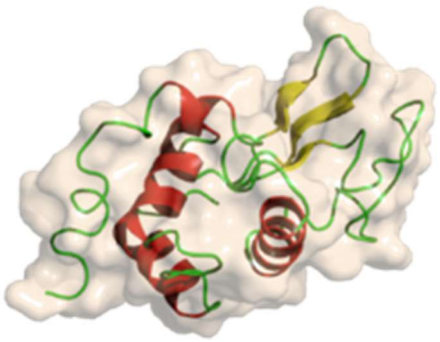
  C'est un test de Fisher exactt (p-value = 0.02155).
  Sous R: fisher.test (matrix(c(17,22,6,0),nrow=2)) p-value = 0.02155
  Si fisher.test(matrix(c(17,22,6,**4**),nrow=2)) p-value = 0.4826
  Il y a 45 sites polymorphes sur cette séquence.
- Que pouvez-vous conclure ?
  surreprésentation des substitutions non synonymes fixées: sélection positive (fixation d'allèles avantageux chez l'une ou l'autre des espèces
- Les auteurs de cette étude (Mc Donald and Kreitman, 1991) ont montré que la séquence codante de l'Adh chez l'ancêtre de *Drosophila simulans* et *Drosophila yakuba* est très proche de la séquence consensus de *Drosophila simulans*. Chez laquelle de ces 2 espèces y-a-t-il eu évolution adaptative suite à la sélection positive ? *Drosophila yakuba*

Partie 3

Le lysozyme est une enzyme bactériolytique présente chez la plupart des mammifères. Chez les primates elle est excrétée dans différentes partie du tube digestif selon le régime alimentaire de l'espèce. Nous étudions une partie de la séquence codante de cette enzyme (les exons 1 et 2) chez **19 espèces représentant les grandes familles de primates (Singes du nouveau monde, Colobinae, Cercopithecinae et Hominoïdes**. Nous cherchons à savoir si cette enzyme a subi une évolution adaptative, et chez quels primates (d'après l'étude de Messier and Stewart, Nature,1997).

Gang de Primates

Lysosyme



- Ouvrir MEGA (v5), et convertissez le fichier «Primates_lysozymes.fasta » en fichier MEGA (.meg). Puis ouvrez ce nouveau fichier (Open a file).
- Construisez un arbre phylogénétique de ces séquences avec la méthode de Neighbor-Joining avec la distance Kimura-2p. Combien de « familles » (ou groupes) observez-vous ? 4
- A quoi correspondent les longueurs de branches ? vitesse d'évolution (nombre de mutations par site)
- A quels types de mutations s'intéresse-t-on ?
  Synonymes / non synonymes
- Quelle approche basée sur l'étude de la divergence entre espèces pourriez-vous utiliser afin d'identifier une éventuelle évolution adaptative du lysosyme ? Que concluez-vous ?
  Approche « phylogénétique » : comparer 2 à 2 les séquences de lysozyme chez toutes les espèces. Pour toutes les comparaisons, on calcule :
  d_N = nombre de subst non synonymes / site non synonymes
  d_S = nombre de subst synonymes / site synonymes
  Et on construit la statistique $Z = (d_N - d_S) / SQRT(Var(d_S) + Var(d_N))$. On a une p-valeur pour chaque comparaison

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
[1]		[-1.016]	[-1.016]	[1.005]	[-0.299]	[0.838]	[0.838]	[1.085]	[1.085]	[1.085]	[1.085]	[0.331]	[0.331]	[0.331]	[0.331]	[0.331]	[0.331]	[-1.151]	[-1.132]
[2]	0.312		[0.000]	[0.007]	[-0.299]	[0.046]	[0.046]	[0.267]	[0.267]	[0.267]	[0.267]	[-0.394]	[-0.394]	[-0.394]	[-0.394]	[-0.394]	[-0.394]	[-1.253]	[-1.233]
[3]	0.312	1.000		[0.007]	[-0.299]	[0.046]	[0.046]	[0.267]	[0.267]	[0.267]	[0.267]	[-0.394]	[-0.394]	[-0.394]	[-0.394]	[-0.394]	[-0.394]	[-1.253]	[-1.233]
[4]	0.317	0.995	0.995		[0.706]	[1.046]	[1.046]	[1.280]	[1.280]	[1.280]	[1.280]	[0.575]	[0.575]	[0.575]	[0.575]	[0.575]	[0.575]	[-1.356]	[-1.340]
[5]	0.766	0.766	0.766	0.481		[0.018]	[0.018]	[0.237]	[0.237]	[0.237]	[0.237]	[-0.415]	[-0.415]	[-0.415]	[-0.415]	[-0.415]	[-0.415]	[-1.277]	[-1.085]
[6]	0.403	0.964	0.964	0.298	0.985		[0.000]	[1.006]	[1.006]	[1.006]	[1.006]	[2.045]	[2.045]	[2.045]	[2.045]	[2.045]	[2.045]	[0.029]	[-0.592]
[7]	0.403	0.964	0.964	0.298	0.985	1.000		[1.006]	[1.006]	[1.006]	[1.006]	[2.045]	[2.045]	[2.045]	[2.045]	[2.045]	[2.045]	[0.029]	[-0.592]
[8]	0.280	0.790	0.790	0.203	0.813	0.317	0.317		[0.000]	[1.430]	[1.430]	[2.300]	[2.300]	[2.300]	[2.300]	[2.300]	[2.300]	[0.216]	[-0.410]
[9]	0.280	0.790	0.790	0.203	0.813	0.317	0.317	1.000		[1.430]	[1.430]	[2.300]	[2.300]	[2.300]	[2.300]	[2.300]	[2.300]	[0.216]	[-0.410]
[10]	0.280	0.790	0.790	0.203	0.813	0.317	0.317	0.155	0.155		[0.000]	[2.300]	[2.300]	[2.300]	[2.300]	[2.300]	[2.300]	[0.216]	[-0.410]
[11]	0.280	0.790	0.790	0.203	0.813	0.317	0.317	0.155	0.155	1.000		[2.300]	[2.300]	[2.300]	[2.300]	[2.300]	[2.300]	[0.216]	[-0.410]
[12]	0.741	0.694	0.694	0.566	0.679	0.043	0.043	0.023	0.023	0.023	0.023		[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[-0.321]	[-0.932]
[13]	0.741	0.694	0.694	0.566	0.679	0.043	0.043	0.023	0.023	0.023	0.023	1.000		[0.000]	[0.000]	[0.000]	[0.000]	[-0.321]	[-0.932]
[14]	0.741	0.694	0.694	0.566	0.679	0.043	0.043	0.023	0.023	0.023	0.023	1.000	1.000		[0.000]	[0.000]	[0.000]	[-0.321]	[-0.932]
[15]	0.741	0.694	0.694	0.566	0.679	0.043	0.043	0.023	0.023	0.023	0.023	1.000	1.000	1.000		[0.000]	[0.000]	[-0.321]	[-0.932]
[16]	0.741	0.694	0.694	0.566	0.679	0.043	0.043	0.023	0.023	0.023	0.023	1.000	1.000	1.000	1.000		[0.000]	[-0.321]	[-0.932]
[17]	0.741	0.694	0.694	0.566	0.679	0.043	0.043	0.023	0.023	0.023	0.023	1.000	1.000	1.000	1.000	1.000		[-0.321]	[-0.932]
[18]	0.252	0.213	0.213	0.178	0.204	0.977	0.977	0.829	0.829	0.829	0.829	0.749	0.749	0.749	0.749	0.749	0.749		[-0.664]
[19]	0.260	0.220	0.220	0.183	0.280	0.555	0.555	0.682	0.682	0.682	0.682	0.353	0.353	0.353	0.353	0.353	0.353	0.508	

Les comparaisons impliquant les espèces de Colobinae et Cercopithecinae, suggèrent H1 dN ≠dS. En test unilatéral on voit que dN > dS. Le nombre de substitutions non synonymes est significativement supérieur au nombre de substitutions synonyme quand on compare ces 2 groupes de primates. Cela traduit une sélection positive forte sur certains résidus d'acides aminés (ici quand on regarde les séquences on constate que 3 acides aminés ont été substitués chez les Clobinae, par rapport aux autres espèces de primates, et surtout par rapport aux Cercopithecinae qui sont les plus proches phylogénétiquement). Le fait que les Cercopthecinae et les colobinae sont proches met en exergue cette signature de la sélection, qui est diluée quand on compare les Colobinae avec des groupes de primates plus distants.

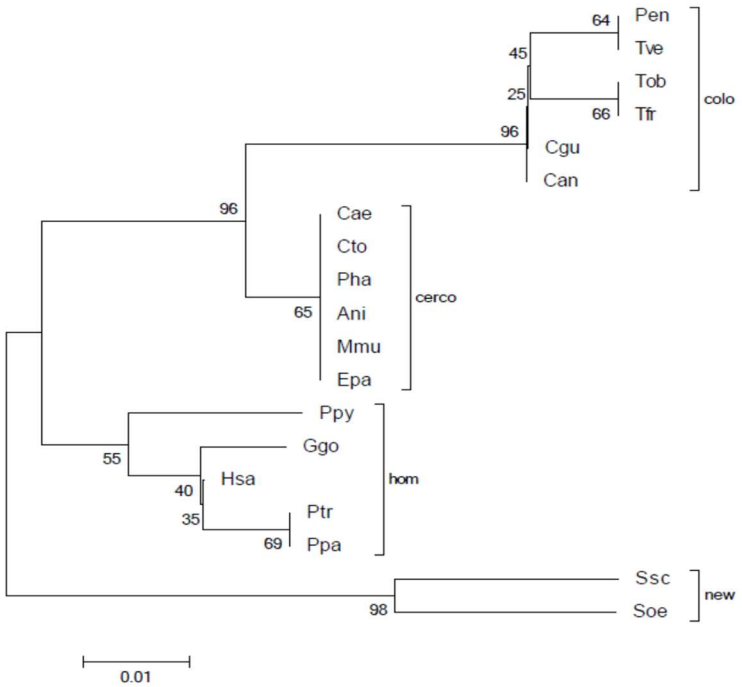
Supplémentaire si assez de temps :

- Pour détecter une adaptation à l'intérieur de chaque groupe (une espèce ou un groupe d'espèce): comparer le d_N - d_S moyen au sein des 4 groupes (hom : pval=1 ; colo : pval=0.064 ; cerco: pval=1 ; new: pval=1. Avec test unilatéral H1=dN>dS)
- Pour détecter une adaptation d'un groupe par rapport à un autre (chez l'ancêtre commun de l'un ou l'autre des groupes) : comparer le d_N - d_S moyen pour des groupes composites tels que « colo/cerco », «colo/hom ». (colo_cerco : pval=0.006 (distances variables sur les paires de séquence) ; colo_hom : pval=1 (distances équivalents sur les paires de séquence); Avec test unilatéral H1=dN>dS

  Pour cela vous devez créer de nouveaux groupes dans l'ongle « Data => Select Taxa and Groups »)

➢ En quoi l'arbre phylogénétique peut vous donner une indication sur le groupe de primates dans lequel le lysosyme a subi une évolution adaptative ?

La branche à la base des Colobinae est plus longue que la branche à la base des Cercopithecinae, à partir de l'ancêtre commun des 2 groupes : dN > dS sur cette branche



➢ Les colobinae (« leaf-eating monkeys ») ont un estomac antérieur complexe dans lequel les bactéries fermentent le matériel végétal, suivi d'un vrai estomac qui exprime des niveaux importants de lysosyme. Quelle interprétation biologique pouvez-vous faire de vos résultats ?

L'enzyme lyzosyme a subi une pression de sélection chez l'ancêtre des colobinae qui avait un régime alimentaire de plus en plus riche en feuilles. Cette pression de sélection était en faveur d'une meilleure dégradation des bactéries, et donc probablement de changements de résidus d'acides aminés améliorant cette dégradation, au niveau du site catalysant l'hydrolyse des glycosaminoglycanes de la paroi des bactéries.

Liste des espèces

Hom : Hsa, Ggo, Ppy, Ptr, Ppa

(*Homo sapiens, Gorilla gorilla, Pongo pygmaeus, Pan troglodytes, Pan paniscus*)
New: Ssc, Soe
(Saimiri sciurus, Saguinus oedipus)
Colo: Pen, Tve, Tob, Tfr, Can, Cgu
(Semnopithecus entellus, Semnopithecus vetullus, Trachypithecus obscurus, Trachypithecus francoisi, Colobus angolensis, Colobus guereza)
Cerco: Cae, Cto, Epa, Ani, Pha, Mmu
(Cercopithecus aethiops, Cercocebus torcatus, Erythrocebus patas, Allenopithecus nigroviridis, Papio hamadryas, Macaca mulatta)