M1 MABS
Examen terminal d'Evolution Moléculaire (durée 2h)
(EM7BMAAME2) – Janvier 2012

Question de cours (2 points)

- Donner la définition d'un profil génétique. Comment l'établit-on ? Pourquoi établir la similarité entre les profils génétiques de plusieurs espèces?
- Comment en pratique identifie-t-on les gènes orthologues à partir des séquences ?

Problème 1 (6 points)

Une partie de l'alignement de la séquence nucléotidique de l'exon 2 d'un gène de 8 individus de *Phytophthora infestans* (pathogène de la pomme de terre) est présenté en Figure 1a. La séquence ancestrale la plus proche a été inférée par une approche phylogénétique, incluant d'autres espèces de *Phytophthora* (Figure 1b). Les sites variables sont indiqués en gras et soulignés. Justifiez à chaque fois vos réponses:

- expliquez quelle(s) force(s) évolutive(s) est/sont à l'origine du profil de polymorphisme observé sur cette séquence. **4 mutations depuis la séquence ancestrale. Sélection positive sur la première mutation (changement d'acide aminé et persque « fixation » dans l'espèce. C'est un balayage sélectif (selective sweep) car la substitution non synonyme (donc à priori neutre) en position 6 est « entraînée » (effet hitch-hiking, auto-stop) du fait de sa proximité au site sélectionné. Le selective sweep aboutit à un excès de variants fréquents (position 6 par exemple) mais aussi un excès d'allèle rares (positions 15 et 27) proche du site sélectionné. On note aussi qu'au niveau de l'haplotype sélectionné, la recombinaison entre positions 15 et 27 a généré des variants haplotypiques. Au plus on s'éloigne du site sélectionné, au plus l'effet de la sélection sur le poymorphisme (le selective sweep) s'affaiblit à cause de la recombinaison.**

Par des approches *in silico*, il a été mis en évidence que cette séquence contient un motif moléculaire pouvant être impliqué dans des interactions protéines-protéines. Une analyse cytologique de plantes infectées par les différents individus de *Phytophthora infestans* indique que (i) la protéine codée par l'individu 8, interagit avec une protéine cytosolique végétale ii) les protéines codées par les individus 1 à 7 sont adressées au noyau de la cellule végétale. De plus, les individus 1 à 7 causent des symptômes nettement plus importants que l'individu 8 sur des plants de pomme de terre.

- donnez une explication biologique évolutive à l'ensemble de ces données.

Chez *Phytophthora infestans*, il semble il y a avoir une corrélation entre le fait de porter la mutation en position 1 et la virulence. Par ailleurs la protéine codée n'interagit pas avec la protéine cytosolique de la cellule végétale, et semble se diriger vers le noyau. Probablement que la mutation a permis le contournement de la « détection » intracellulaire du parasite par la plante et que cette protéine peut du coup exercer une fonction (associée peut être à un autre domaine) sur les gènes de la plante (peut être une inactivation de certains gènes de défense). Le parasite a donc subi une pression de sélection par la plante afin de ne pas être reconnu. La mutation étant bénéfique pour le

parasite, elle a été sélectionnée positivement (augmentation de sa fréquence dans l'espèce).

Figure 1a: séquence de l'exon 2, chez 8 individus de *Phytophthora infestans*:

```
1 - ACT GCG CTG TCC CCG CGC TCA GGC AGC
2 - ACT GCG CTG TCC CCA CGC TCA GGC AGT
3 - ACT GCG CTG TCC CCG CGC TCA GGC AGT
4 - ACT GCG CTG TCC CCA CGC TCA GGC AGC
5 - ACT GCG CTG TCC CCA CGC TCA GGC AGC
6 - ACT GCG CTG TCC CCA CGC TCA GGC AGC
7 - ACT GCT CTG TCC CCA CGC TCA GGC AGC
8 - TCT GCT CTG TCC CCA CGC TCA GGC AGC
```

Figure 1b: séquence de l'exon 2, chez l'ancêtre le plus proche de *Phytophthora infestans*:

```
TCT GCT CTG TCC CCA CGC TCA GGC AGC
```

```
NB:
ACT = Thr, TCT = Ser,
GCG = Ala, GCT = Ala,
CCA = Pro, CCG = Pro
AGC = Ser, AGT = Ser,
```

Problème 2 (5 points)

Zhenguo et collaborateurs (Nucleic Acids Res., 2007, 1-13) ont conduit une analyse phylogénétique des composants clés du système de réparations des mésappariements au niveau de l'ADN. Ces mésappariements sont introduits par l'ADN polymérase lors de la division cellulaire et leur non correction entraînerait des mutations. Le gène codant pour la protéine MutS est un acteur de ce système de réparation. Chez *Escherichia coli* il a été démontré qu'un homodimère de MutS se lie sur les nucléotides mal appariés et forme un complexe MutS-ADN. Cependant plusieurs gènes homologues à *mutS* existent dans *E. coli* et une recherche de similarité dans 461 génomes bactériens a montré qu'on pouvait distinguer 4 sous-familles de protéines MutS, MutS1 (correspondant à la protéine MutS dont la fonction est décrite ci-dessus) à MutS4. Chez les eucaryotes, 7 gènes homologues aux gènes MutS bactériens ont été identifiés. Les arbres phylogénétiques ont montré qu'ils appartenaient aux sous-familles MutS1 et MutS2. Nous allons nous intéresser ici à l'évolution des séquences protéiques MutS2. Ces dernières n'ont été identifiées que dans les génomes nucléaires des espèces possédant des chloroplastes, à savoir plantes à fleurs et algues vertes. L'analyse évolutive de cette sous-famille de protéines a été réalisée sur un ensemble d'espèces représentatives et a conduit à l'arbre de la figure 2. La reconstruction phylogénétique a été réalisée à l'aide de deux méthodes la Neighbor Joining (NJ) et maximum de vraisemblance (PhyML). Les deux topologies étant congruentes, *i.e.* les mêmes, seul l'arbre obtenu par la NJ est présenté mais les valeurs de bootstrap obtenues avec chacune des méthodes sont reportées (NJ et PhyML).

- Analyser l'arbre de la figure 2 pour émettre des hypothèses quant à l'origine et l'évolution de ce gène : duplication, transferts horizontaux, endosymbiose (il est actuellement accepté que les mitochondries proviendraient d'endosymbiontes

apparentées aux α-protéobactéries et les chloroplastes d'endosymbiontes apparentées aux cyanobactéries), etc.

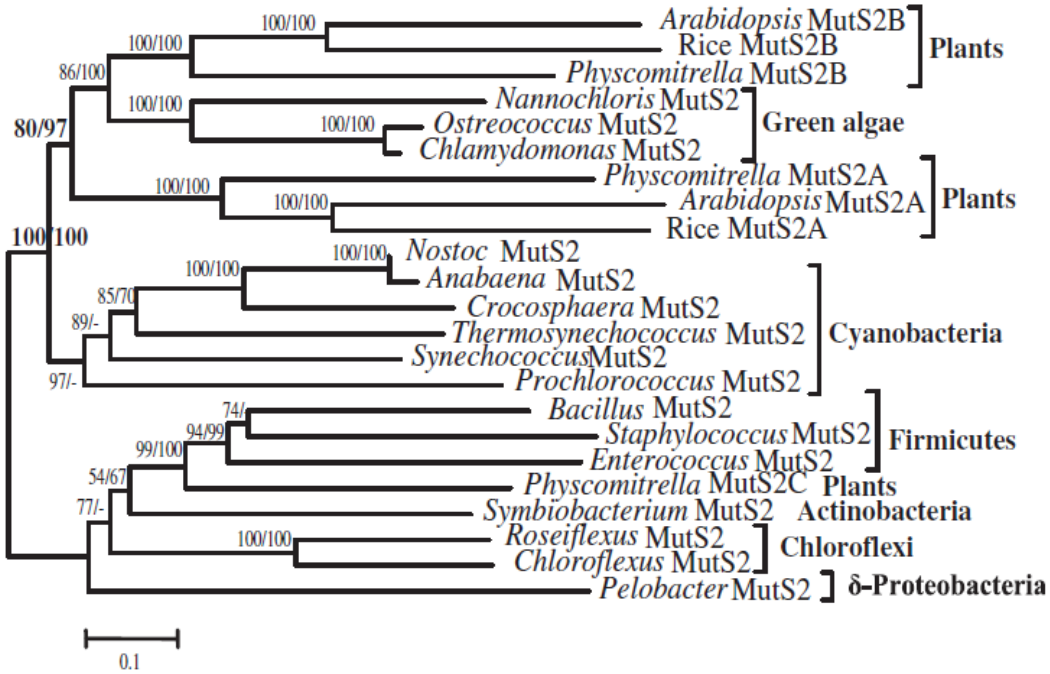

Figure 2: Arbre phylogénétique obtenu sur un ensemble de séquences protéiques de la sous-famille MutS2 (extrait de Zhenguo *et al.*, 2007, Nucleic Acids Res., 1-13).

Les génomes nucléaires eucaryotes de plantes possèdent deux copies du gène *mutS2* correspondant aux deux protéines MutS2A et MutS2B. Le génome de *Physcomitrella* possède une troisième copie de ce gène correspondant à la protéine MutS2C. Les génomes procaryotes ne possèdent qu'un seul exemplaire du gène. Le sous-arbre correspondant aux séquences eucaryotes (plantes et algues vertes) possède un ancêtre commun avec le sous-arbre correspondant aux séquences de cyanobactéries. La branche conduisant à ce nœud ancêtre est très bien supportée (valeur de bootstrap de 100). La topologie de l'arbre est donc en accord avec l'hypothèse actuellement bien acceptée que les chloroplastes aient dérivés d'un endosymbionte ancestral apparenté aux cyanobactéries. Le gène eucaryote *mutS2* aurait donc tout d'abord été acquis par endosymbiose, ce qui explique qu'il ne soit trouvé que dans les génomes de plantes possédant des chloroplastes. Il aurait ensuite été transféré du génome ancestral chloroplastique au génome nucléaire. De plus, le gène *mutS2* aurait été dupliqué et aurait donné naissance à deux gènes paralogues *mutS2A* et *mutS2B* avant la divergence entre plantes terrestres et algues vertes car les deux protéines MutS2A et MutS2B forment deux sous-arbres distincts. Le génome ancestral des espèces d'algues vertes auraient perdues la copie *mutS2A* car un seul gène est identifié dans ces organismes et son produit protéique MutS2 forme un cluster sister groupe de celui regroupant les protéines MutSB (les deux groupes possèdent un ancêtre commun).
Le positionnement de la séquence MutS2C de *Physcomitrella* en groupe externe des séquences MutS2 de Firmicutes suggère que le gène codant pour cette protéine ait été acquis par transfert horizontal probablement à partir d'un génome de Firmicutes.

Problème 3 (7 points)

Afin de rechercher des gènes fonctionnellement importants dans le génome de la plante modèle *Medicago truncatula*, des chercheurs ont recherché des traces de sélection naturelle chez deux lignées vivant dans des environnements très distincts. Un test de neutralité

de Tajima a été effectué sur 30 gènes dans chacune des deux lignées, basé sur l'alignement des séquences de plusieurs individus pour chaque lignée. La formule de la statistique *D* du test de Tajima est :

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{SE(\hat{\theta}_\pi - \hat{\theta}_S)}$$

où θ_π et θ_s sont deux estimateurs de la diversité au niveau nucléotidique sur une séquence (SE = ecart-type).

- Lequel de ces deux estimateurs évalue la probabilité qu'un nucléotide soit polymorphe? $\boldsymbol{\theta_s}$ .Expliquez comment vous feriez pour le calculer. **Sur l'alignement, on compte le nombre de sites polymorphes, et on divise par la longueur de la séquence (il y a un facteur de correction pour la taille de l'échantillon, mais ce n'est pas très grave s'ils ne le mentionnent pas) ; ce qui donne une proba de polymorphisme par paire de base.**
- Lequel de ces deux estimateurs évalue la probabilité d'hétérozygotie par nucléotide? $\boldsymbol{\theta_\pi}$. Expliquez comment vous feriez pour le calculer. **Sur l'alignement, on compte le nombre de différences nucléotidiques par paires de séquences; on fait la moyenne sur l'ensemble des paires, et on divise par la longueur de la séquence; ce qui donne une proba d'hétérozygotie par paire de base.**

Le tableau ci-dessous résume les résultats du test de Tajima. Les 30 gènes sont classés selon les différents types de valeurs prises par *D* et selon les mécanismes moléculaires auxquels ils sont associés.

Lignée	Mécanismes moléculaires	$D > 0$	$D < 0$	$D = 0$
1	stress abiotique prolongé	0	3	10
	résistance aux pathogènes (gènes paralogues)	12	0	5
2	stress abiotique prolongé	0	10	3
	résistance aux pathogènes (gènes paralogues)	5		12

- Que signifient des valeurs de $D = 0$, < 0 et > 0.
  - **D=0 : $\theta_\pi = \theta_s$, ce qu'on observe sous neutralité**
  - **D>0 : $\theta_\pi > \theta_s$, le taux d'hétérozygotie est supérieur : beaucoup de variants de fréquence intermédiaire qui augmente l'hétérozygotie, pour une même proba de polymorphisme : Sélection balancée (maintient des plusieurs allèles dans la pop)**
  - **D<0 : $\theta_\pi < \theta_s$, le taux d'hétérozygotie est inférieur: beaucoup de variants de fréquence faible (allèles rares) qui influencent peu l'hétérozygotie. Sélection purifiante (background sélection) = élimination des allèles délétères (excès d'allèles rares. Le balayage sélectif (sélection positive très forte) peut aussi occasionner un excès d'allèles rares sur les sites nucléotidiques environnants (en même temps qu'un excès d'allèles fortement fréquents).**

D est sensible à la fréquence des variants.

- Interprétez les différences entre lignées en termes de pressions de sélections exercées par leur environnement et de conséquence sur le polymorphisme moléculaire de ces gènes (justifiez qualitativement).

  NB : le test est significatif au seuil 0.001 pour tous les gènes avec $D > 0$ et $D < 0$

  La lignée 1 : beaucoup de gènes de résistance sont sous pression de sélection balancée. Peu de gènes de stress abiotique présentant une sélection purifiante (ou balayage sélectif)
  La lignée 2 : beaucoup de gènes de stress abiotique présentant plutôt un balayage sélectif (la sélection purifiante est plutôt une sélection sur le très long terme et qui est focalisée sur certains gènes ou parties de gènes). Peu de gènes de gènes de résistance sont sous pression de sélection balancée.

  La lignée 1 de Medicago semble évoluer dans un environnement contenant une forte diversité de pathogènes qui lui oblige à présenter un répertoire allélique important au niveau des gènes de résistances (souvent lié à la perception de molécules microbiennes), d'où une sélection pour le polymorphisme. Cependant le milieu n'est pas limitant en terme de ressources abiotiques (ex : accès à l'eau, sel, N, lumière) : relâchement de la pression de sélection pour des gènes du stress abiotique prolongé.

  Lignée 2 : sélection positive pour améliorer l'accès aux ressources et la gestion de ces ressources, dans un environnement carencé. Mais environnement peu riche en pathogènes.

  On peut imaginer :
  Lignée 1 : peut être milieu riche, humide, forte diversité biologique de pathogènes
  Lignée 2 : milieu pauvre (ex : désertique), faible diversité biologique de pathogènes