TD N°3 — LA PHYLOGÉNIE DES γ-PROTÉOBACTÉRIES : INFÉRENCE BAYÉSIENNE ET CORROBORATION MULTIGÉNIQUE.

Objectif : Utilisation du logiciel d'inférence bayésienne **MrBayes**, version 3.2, de John Huelsenbeck & Frederik Ronquist [http://mrbayes.sourceforge.net/].

Modèle biologique : la phylogénie des γ-protéobactéries sera étudiée à l'aide de séquences d'acides aminés déduites *in silico* des génomes d'*Escherichia coli* K12, *Buchnera aphidicola* APS, *Haemophilus influenzae* Rd, *Pasteurella multocida* Pm70, *Salmonella typhimurium* LT2, *Yersinia pestis* CO_92, *Yersinia pestis* KIM5 P12, *Vibrio cholerae*, *Pseudomonas aeruginosa* PA01, et *Wigglesworthia glossinidia*. Trois groupes externes seront utilisés : *Xanthomonas axonopodis* pv. citri 306, *Xanthomonas campestris*, et *Xylella fastidiosa* 9a5c. Quatre protéines seront ici étudiées : atpB, bioB, gyrA, et rpoA [Article de référence : Lerat E, Daubin V, Moran NA (2003) From gene trees to organismal phylogeny in Prokaryotes: The case of the gamma-Proteobacteria. PLoS Biology 1:101-109 (PDF sur le site http://www.plosbiology.org/)].

Reconstruction d'arbres sous **MrBayes**. Les commandes à utiliser pour inférer l'histoire évolutive des protéines atpB, bioB, gyrA, et rpoA sont les suivantes.

Begin mrbayes;			Initie les commandes.
	PRSET		"*Priors set*" : déclare les distributions *a priori*.
		AAMODELPR=FIXED(WAG**);**	Spécifie le modèle WAG (**AA** : amino-acides).
	LSET		"*Likelihood set*" : déclare d'autres paramètres du modèle.
		RATES = GAMMA	Spécifie l'hétérogénéité des taux entre sites par la loi **Γ**.
		NGAMMACAT = 4;	Nombre de catégories de la loi Gamma (**Γ**) discrète.
	MCMC		Démarre l'inférence par MCMC.
		NCHAINS = 4	Nombre de chaînes MCMC lancées pour l'exploration de l'espace des paramètres.
		NGEN = 1000	Nombre total de **générations** MCMC (**G**).
		PRINTFREQ = 10	Fréquence d'affichage à l'écran (en **nombre de générations**).
		SAMPLEFREQ = 10	Fréquence d'échantillonnage (**S**), avec **G/S** échantillons retenus.
		NRUNS = 2	Nombre d'analyses lancées en parallèle.
		DIAGNFREQ = 100;	Fréquence de calcul du diagnostic de convergence entre les **NRUNS** = 2 analyses parallèles.
	SUMP	**BURNIN** = A;	"*SUMmarize Parameters*" : calcul des **paramètres** sur **G/S — A** échantillons, avec un allumage ("*burnin*") de **A** échantillons.
	SUMT	**BURNIN** = A;	"*SUMmarize Trees*" : calcul des **arbres** sur **G/S — A** échantillons, avec un allumage de **A** échantillons.
End;			Clot les commandes.

En reportant graphiquement l'évolution du logarithme de vraisemblance (ln*L*) et de la moyenne des fréquences des bipartitions ("*ASDSF*") en fonction du nombre de générations **G**, identifiez la durée de la phase d'allumage (soit un total de **A** x **S** générations).

Inférez les arbres les plus probables *a posteriori* pour chacune des 4 protéines. Calculez aussi les paramètres du modèle (α de loi **Γ**, et longueur de l'arbre). Déduisez-en les taux relatifs de remplacement des acides aminés dans les 4 protéines.

Comparez leurs topologies, ainsi que les probabilités postérieures (**PP**) de chacun de leurs clades. Comment expliquez vous l'intégralité des résultats observés ?