bee@lyon
BIODIVERSITÉ, ÉCOLOGIE, ÉVOLUTION

# STAGE DE RECHERCHE M2 ECOLOGIE EVOLUTION GENOMIQUE
## Rentrée 2017

---

## Etude évolutive de la composition des génomes bactériens : le rôle des processus non adaptatifs

Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558 Villeurbanne
Encadrants : Vincent Daubin (vincent.daubin@univ-lyon1.fr), Laurent Guéguen (laurent.gueguen@univ-lyon1.fr)

**Contexte**

Au sein de l'ensemble des espèces bactériennes et archées, la composition des génomes est extrêmement variable. En considérant une statistique simple telle que la composition en bases G+C des génomes, on observe que ce contenu en G+C va de 15 % à 75 %. Mais aussi, de façon plus intrigante, on observe que le contenu moyen en G+C de 50 % a tendance à être évité.

Ces variations sont dues à l'action conjointe de processus non-adaptatifs (mutation, dérive génétique, conversion génique biaisée) et adaptatifs (sélection sur les séquences protéiques, sur les codons préférés, voire sur d'autres propriétés des séquences), mais le rôle relatif de ces pressions évolutives est encore le sujet de débats brûlants.

Comprendre le rôle exact de ces différents mécanismes, et relier leur action à l'écologie des espèces est l'une des questions clés en génomique bactérienne.

**Objectifs du stage**

L'objectif de ce stage est de trouver une explication évolutive à cette variation de composition.

Pendant plus de 50 ans, on a pensé que l'ampleur de cette variation était due essentiellement aux différences de mutations entre génomes (certains mutant plutôt vers les nucléotides A et T, d'autres plutôt vers les nucléotides G et C).

Mais l'analyse fine des patrons de mutations sur un grand nombre de génomes complets suggère que les mutations poussent toujours dans la direction d'une composition plus forte en A et T. La question se pose donc de comprendre quelles forces évolutives maintiennent des génomes ayant des compositions en G et C moyennes ou fortes.

Pour apporter un nouveau regard sur cette question, nous avons développé un modèle d'évolution des séquences codantes (dénommé SENCA), qui distingue les niveaux auxquels agissent les différents processus influent sur la composition en GC. En particulier, il permet de distinguer les pressions agissant à trois niveaux pertinents : nucléotidique, usage des codons synonymes et usage des acides aminés. La décomposition en ces différents niveaux permet de mieux disséquer les pressions de mutations et de sélection qui s'exercent sur les séquences.

Les études préliminaires que nous avons effectuées grâce à ce modèle sur des génomes bactériens montrent que les rôles respectifs de ces mécanismes sont très variables, et dépendent de la profondeur phylogénétique (Fig 1).

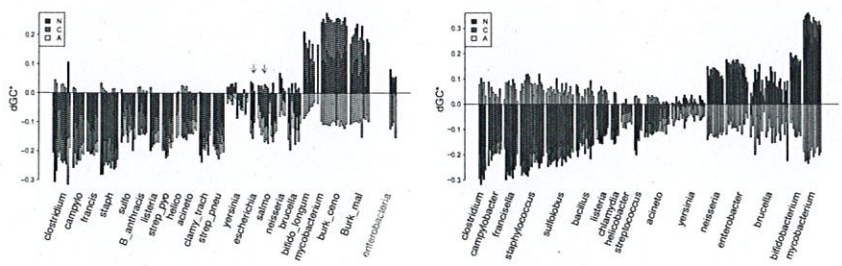


Figure 1 - Décomposition en couches des processus évolutifs au sein de plusieurs espèces bactériennes. Ces espèces sont rangées par G+C croissant. Description des couches : **N** : mutation, **C** sélection sur l'usage des codons synonymes, **A** sélection de l'usage des acides aminés. Gauche : analyses intra-spécifiques. Droite : analyses inter-spécifiques.

Nous disposons, sur de multiples espèces de bactéries et archées, de grands ensembles de séquences de gènes alignées, et prêtes pour une étude phylogénétique approfondie.

Le travail consistera à utiliser le modèle SENCA pour déconstruire les processus évolutifs qui expliquent ces données. Cette analyse permettra de confirmer ou d'infirmer l'hypothèse du biais mutationnel universel vers A et T et de comprendre quels sont les processus qui s'y opposent.

**Profil Requis**

1. Maîtrise de langages d'analyse et de scripts (R, shell ou python) ;
2. Génétique, évolution, phylogénie.

**Bibliographie associée**

1. Pouyet F., Bailly-Bechet M., Mouchiroud D., Gu_eguen L. (2016) SENCA : a multilayered codon model to study the origins and dynamics of codon usage, GBE, doi:10.1093/gbe/evw165.