Graines espacées et recherche d'ARN non-codants

Les **ARN non-codants (ARNnc)** ont des fonctions multiples et essentierlles dans la cellule. Pour les identifier *ab initio*, on recourt à des outils de **comparaison de séquences** : la génomique comparative sert de filtre à une recherche de structure secondaire. Est-il possible d'utiliser des méthodes plus sensibles que Blast afin de mieux explorer les génomes à la recherche d'ARNnc ?

Recherche de séquences conservées : les graines espacées

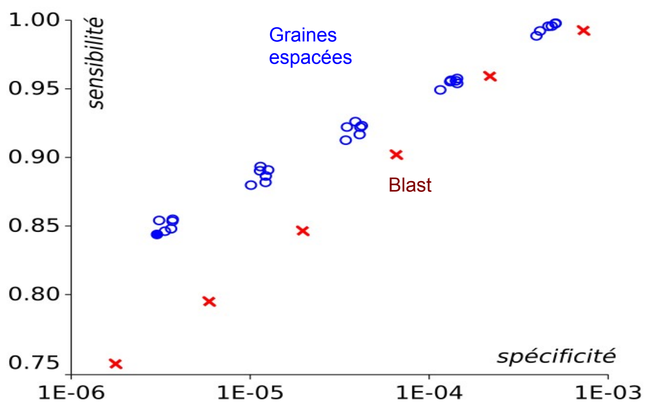
Les programmes tels que Fasta ou Blast sont des **heuristiques**. Ils recherchent tout d'abord des *k*-mers entièrement conservés, les **graines contiguës**, puis étendent ces graines pour construire des alignements.

A même efficacité et spécificité, les **graines espacées** apportent une **meilleure sensibilité**, permettant ainsi de trouver des régions moins conservées [Ma2002].

La graine contiguë #### ne détecte que le premier alignement, alors que la graine espacée ##-## détecte les deux.

```
GACTGAACTCAT   TAGACTCGACGA
|.||.||||.||   |||.||.|||.|
GGCTAAACTAAT   TAGGCTAGACTA

    ####
  ##-##          ##-##
```



Compromis spécificité-sensibilité sur des alignements (plus de 50% d'identité) provenant de 574 familles de la base de données d'ARN non-codants RFAM.

Sensibilités sur des benchmarks de [Freyhult 2007]

E-valeur : 10^{-4}

	ARNt 1114 séquences	ARNnc U5 235 séquences	ARNr 5S 602 séquences
Blast	0.04	0.85	0.32
Yass	0.18	0.93	0.59

En général, les ARNnc sont **moins conservés** que les régions codantes, donc plus difficiles à déceler par génomique comparative.

Pour évaluer le gain pratique que peuvent apporter les algorithmes à base de graines espacées, nous avons comparé systématiquement les résultats obtenus avec Blast [Altschul1991] et avec Yass [Noé2005].

Les deux expériences font apparaître de **meilleurs résultats** avec Yass qu'avec Blast.

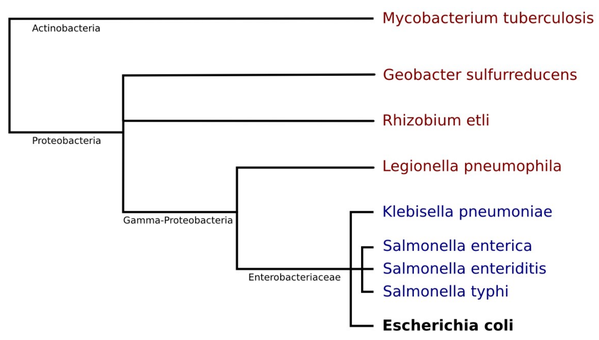
génomique comparative — graines espacées — ARN non-codants

Annotation de génomes

Pour étudier le gain des techniques à base de graines espacées **à l'échelle de génomes bactériens,** nous avons repris les travaux de [Rivas2001] pour une recherche de candidats ARNnc chez *E. coli*.

séquence pivot
Zones intergéniques
E. coli (550 Kb)

4 génomes bactériens, proches ou éloignés



Comparaison en étoile

Sélection de zones d'intérêt

~ 1000 zones

Nombre de ARNnc contenus ou intersectés

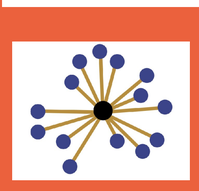
E-valeur théorique : 10^{-2}

	160 ARNnc Blast	160 ARNnc Yass
espèces proches	153	155
espèces éloignées	109	153

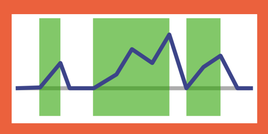
Mise en place d'un pipeline d'annotation

http://bioinfo.lifl.fr/

Génomes complets



Comparaisons (Yass)



Sélection de zones d'intérêt

Découverte de candidats ncRNAs
Recherche de structure secondaire (caRNAc)

Arnaud Fontaine, Mathieu Giraud, Laurent Noé, Hélène Touzet
Bioinfo/Sequoia, LIFL, INIRA, UMR CNRS 8022, Université Lille 1