MOSAIC

(Projet de Recherche 2003)

ACI IMPBIO

Analyse comparative de génomes bactériens

http://www-mig.jouy.inra.fr//aci-mosaic/

CONTEXTE

Chez certaines espèces bactériennes des comparaisons de génomes de différentes souches ont permis de mettre en évidence une **structure en mosaïque** qui est composée d'un **« squelette »** commun à toutes les souches et de **« boucles »** spécifiques de chaque souche. Il est important d'étendre cette approche à toutes les espèces pour lesquelles on dispose de suffisamment d'information et de tenir compte de cette structure hétérogène dans l'analyse des chromosomes bactériens

OBJECTIFS

La plupart des analyses de comparaison de génomes bactériens effectuées jusqu'à présent s'intéressent principalement au contenu en gènes des génomes étudiés. Il est cependant bien clair que l'information génétique ne se réduit pas aux seuls gènes. De nombreux types de motifs d'ADN de fonction connue ont été identifiés dans les génomes bactériens. Au delà de ceux qui ont déjà été identifiés, il est très probable que d'autre motifs existent, qui ne sont pas connus pour le moment. L'analyse de ces motifs représente l'un des défis majeurs de l'étude des génomes aujourd'hui. L'objet de notre étude est de développer des méthodes alliant la comparaison de génomes et l'analyse statistique pour analyser la distribution et prédire des motifs de fonction connue ou potentielle. Ces méthodes sont appliquées à des exemples de motifs biologiques et les prédictions testées expérimentalement.

PROJET

Analyse de la structure mosaïque des génomes

stratégie systématique de segmentation en boucles et squelette des génomes bactériens, base de données permettant d'accéder aux informations de segmentation et de les croiser avec les annotations des génomes, analyse des boucles.

Développements statistiques

analyse de la distance relative entre différents motifs, analyse du comptage des motifs en modèle hétérogène, modèles de mélanges pour l'analyse des boucles

Prédiction de motifs structurant les génomes

identification de sites Chi (impliqués dans la réparation du chromosome) de diverses bactéries. Analyse des séquences KOPS (impliquées dans la ségrégation des chromosomes) chez *Escherichia coli*. Identification de motifs structurant le chromosome d'*E. coli* en macrodomaines.

RESULTATS

•MOSAIC: Base de données de comparaison systématique de génomes, http://genome.jouy.inra.fr/mosaic/ Analyse des boucles chez *E. coli* [1,2]
•Méthode de typologie de boucles par modèles de mélange.
•Modélisation statistique de l'influence entre les occurrences de deux motifs [3]
•Statistique des comptages de mots dans un modèle de Markov caché. Nouvelle loi de poisson composée pour approcher la loi du comptage d'une famille de mots rares [4]
•Statistique de comparaison d'exceptionnalité pour un motif entre deux séquences.
•Prédiction *in silico* du site Chi de *Staphylococcus aureus*. Confirmation expérimentale
•Analyse de la distribution des motifs KOPS [5]
•Stratégie de prédiction de motifs spécifiques de domaines

CALENDRIER

2003	(novembre) Démarrage
2004	(mai) **Réunion générale,**(juin, juillet, septembre, octobre, décembre) Réunions de travail
2005	(février, mai, juin) Réunions de travail,(mars) **Réunion générale** (bilan à mi parcours);(mai , juin) Réunions de travail

LABORATOIRES IMPLIQUES

•Unité de Recherches Laitières et Génétique Appliquée, INRA, Jouy en Josas
•Mathématique Informatique et Génome, INRA, Jouy en Josas
•Département organisation de Modélisation de l'Information et des Processus, INA-PG, Paris
•Centre de Génétique Moléculaire, CNRS, Gif-sur-yvette

MOTS CLES

Structure squelette/boucle, motifs structurant le chromosome (Chi, KOPS), alignement de génomes complets, bases de données, statistiques de motifs, motifs co-répartis, séquences hétérogènes

PUBLICATIONS PRINCIPALES

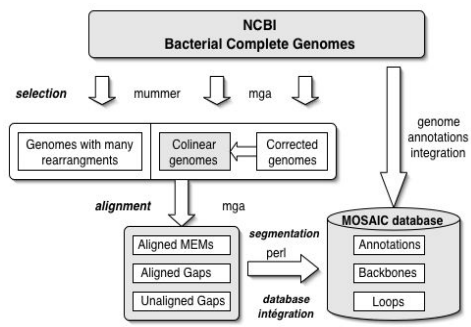
[1] Chiapello, H., Bourgait, I., Sourivong, F., Heuclin, G., Jacquemard, A., Petit, M-A, and El Karoui, M. *Systematic determination of the MOSAIC structure - backbone vs strain specific loops - of bacterial genomes.*(sous presse, BMC Bioinformatics)

[2] Chiapello, H., Bourgait, I., Sourivong, F., Jacquemard, A., Petit, M-A, and El Karoui, M. *MOSAIC : an online database for systematic determination of the mosaic structure of bacterial genomes* JOBIM, Lyon, 2005

[3] Gusto, G. et Schbath, S. FADO : *a statistical method to detect favored and avoided distances between occurrences of motifs using the Hawkes' model* (en revision, Statistical applications in genetics and molecular biology)

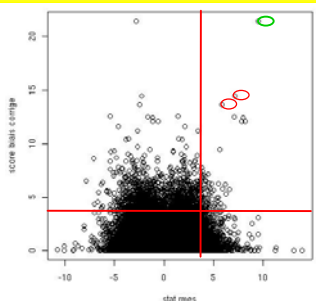
[4] Roquain E. and Schbath S. *a new compound Poisson approxiation for the distribution of the count of a word family in a markovian sequence.* (in preparation)

Base de donnée MOSAIC



Représentation schématique du traitement des données dans la base MOSAIC. Les génomes ne comportant pas de réarrangements sont sélectionnés en utilisant Mummer puis alignés avec MGA. Il sont ensuite segmentés et intégrés dans MOSAIC.

Analyse des séquences Chi et KOPS



L'objectif est d'identifier des motifs sur-représentés (abscisse >3) et significativement plus fréquents sur un brin du chromosome (ordonnée >4). Le motifs identifié en vert correspond au motif Chi. L'analyse a permis d'identifier deux motifs (en rouge) dont l'un au moins stimule l'activité d'une protéine impliquée dans la ségrégation des chromosomes chez *E. coli*.