FouDanGA

(Projet de Recherche 2004)

Fouille de données pour l'annotation de génomes d' actinomycètes

http://www.loria.fr/research/aci/imp/

CONTEXTE

• Mise en œuvre de méthodes de fouilles de données pour comprendre les mécanismes impliqués dans l'expression, la transmission et l'évolution des gènes

• Développement de méthodes combinatoires et stochastiques pour prédire les séquences promotrices et autres petites séquences régulatrices chez les bactéries

• Fouille sur les génomes *S. coelicolor*, *S. avermitilis* et *M. tuberculosis* ainsi que le génome de *S. ambofaciens* en cours de séquençage par l'UMR 1128 en collaboration avec le Génoscope (CNS, Evry).

OBJECTIFS

• Extraction par des méthodes de fouille de données de motifs d'ADN: SFFT comme Site de Fixation de Facteur Transcriptionnel (impliqués dans la régulation de l'expression génique chez les bactéries du groupe des actinomycètes)

• Le groupe des actinomycètes comprend :

des espèces d'intérêt industriel comme les *Streptomyces* (producteurs d'antibiotiques microbiens)

des espèces pathogènes comme certaines mycobactéries (*Mycobacterium tuberculosis)*

• Deux approches informatiques développées :

utilisation d'algorithmes de recherche de mots puis de couples de mots sur-représentés dans les régions en amont de gènes orthologues d'espèces phylogénétiquement proches

fouille de données génomiques sans *a priori* pour faire émerger des sous-séquences d'ADN dans les régions intergéniques (**spécification apprentissage** et **utilisation** de modèles de Markov cachés du second-ordre (HMM2), pour faire apparaître des irrégularités dans des grandes séquences d'ADN)

SIGffRid [4]

Principe de SIGffRid - Recherche des sites promoteurs d'une bactérie B1 en utilisant des informations issues d'une bactérie B2 phylogénétiquement proche de B1. Cette démarche, généralisable à tout couple de bactéries proches, permet de prédire les SFFT qui leur sont communs.

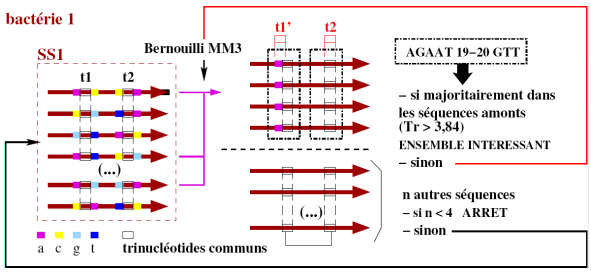
Etapes de l'algorithme

A- **Définition** avec R'MES* **d'un dictionnaire *D* de mots** (de 3 à 7 lettres) **de B1 statistiquement intéressants**.

B- **Détermination d'un lot de paires de séquences orthologues de B1 et B2** (MGBD). **Recherche**, pour chaque paire de séquences amonts d'orthologues s_{1i} et s_{2i} ($i \in [1..n]$), de triplets $C_i=\{w^1_i,w^2_i,\{s_{1i},s_{2i}\}\}$ avec w^1_i et w^2_i des mots conservés appartenant à D séparés par un espacement variable e.

C- **Recherche de paires de trinucléotides dans les C_i** : Pour chaque triplet (t1,t2, d) possible, avec t1 et t2 des trinucléotides et d un espace, en considérant l'ensemble des C_i obtenus, **création d'un ensemble G_j** contenant les C_i qui vérifient: $(t1 \subset w^1_i) \wedge (t2 \subset w^2_i) \wedge (d \in [e..e+1])$.
Pour chaque G_j, **regroupement des séquences** : $SS_{1\,j} = Us_{1i} \in G_j$ et $SS_{2\,j} = Us_{2i} \in G_j$.

D- **Extension des trinucléotides et création des motifs candidats en fonction des séquences des SS_{1j} et de critères probabilistes:**



Regroupement et extension des motifs candidats dirigés par approche statistique. T_R correspond à la mesure de la signi-ficativité de la spé-cificité du motif pour les séquences amonts.

Résultats

3 SFFT connus sont retrouvés ou confirmés, avec un grand nombre de nouveaux gènes co-régulés candidats pour chacun.

Au moins deux nouveaux SFFT sont proposés, à la fois chez *Streptomyces coelicolor* et *Streptomyces avermitilis*.

LABORATOIRES IMPLIQUES

LORIA (UMR7503), INRIA Lorraine, Laboratoire de Génétique et Microbiologie (UMR INRA 1128, UHP Nancy)

MOTS CLES

Fouille de données, approche comparative, SFFT, facteur sigma

PUBLICATIONS PRINCIPALES

Production scientifique

[1] M. Benoit , F. Le Ber, J.-F. Mari, C. Mignolet, C. Schott. CarrotAge, *Un logiciel pour la fouille de données agricoles*, Colloque STIC et Environnement SE'2003, Rouen, 2003.

[2] F. Touzain, I. Debled-Rennesson, B. Aigle, P. Leblond et G. Kucherov.*Identification of Transcription Factor Binding Sites in Streptomyces coelicolor* A3 (2) by Phylogenetic Comparison. Poster, ECCB, Paris, 2003.

[3] C. Eng, A. Thibessard, S. Hergalant, J.-F. Mari, P. Leblond. *Data Mining Using Hidden Markov Models to Detect Heterogeneities into Bacterial Genomes*, poster JOBIM, Lyon, 2005.

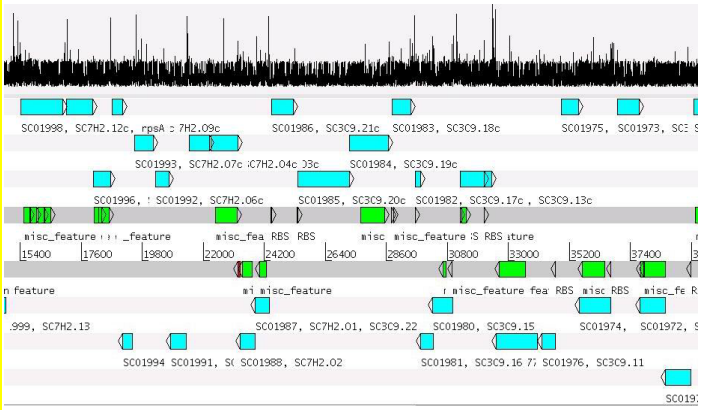
[4] F. Touzain, S. Schbath, I. Debled-Rennesson, B. Aigle, P. Leblond et G. Kucherov. SIGffRid: *Programme de recherche des sites de fixation des facteurs de transcription par approche comparative*. Communication longue JOBIM, 2005.

[5] Hergalant S., Aigle B., Leblond P. et Mari J.-F. *Fouille de données du génome à l'aide de modèles de Markov cachés*. Ateliers EGC 2005, Paris, 141-148.

Référence

[*] S. Schbath. *An efficient statistic software to detect over- and under-represented words in dna sequences*. *J. Comp. Biol.,* 4 :189-192, 1997.

HMM2 [5]



1. Recherche des pics sur 1,2 Mb d'ADN de *S. coelicolor*
2. Classification hiérarchique par alignement multiple (*Clustalin*)
3. Pour chaque classe m de motifs faire
   Définition du consensus de la classe m
   Recherche des occurences de m sur le génome
4. Sélection des paires m1-d-m2 séparées par 0 d 25 nucléotides
5. Recherche des occurences de m1-d-m2 sur le génome (isolats)
6. Sélection des occurences de m1-d-m2 situées majoritairement en proche amont des ORF (<500 pb)

Conclusions

• Spécification de deux méthodes combinatoire / stochastique pour la recherche de SFFT.
• Améliorations conjointes des deux systèmes
• Validations biologiques plus complètes à venir