Proposition de Stage de Master Recherche 2ème année

LIX École polytechnique
Équipe bioinformatique

Responsable du stage: Pierre Nicodème
`nicodeme@lix.polytechnique.fr` tel: 01.69.33.28.64

Analyse statistique de la filtration par hits de Blast

Pour accélérer la recherche d'alignements entre deux séquences, Blast utilise une méthode de points d'ancrage (ou *hits*). Un hit est une paire de mots courts fortement similaires et appartenant l'un à la séquence requête, et l'autre à la séquence cible. La notion de similarité est évaluée en utilisant une matrice de similarité, en général appartenant aux familles BLOSUM ou PAM pour les protéines.

Dans l'approche 1-hit, lorsqu'un hit est trouvé, Blast cherche à l'étendre à droite et à gauche en progressant simultanément sur les deux séquences. Les versions récentes de Blast effectuent ces extensions en autorisant des gaps (insertions-suppressions).

Dans l'approche 2-hits, on ne fait une extension que si l'on a trouvé deux hits dans une bande diagonale du rectangle utilisé pour faire un alignement global des deux séquences.

La filtration utilise un score minimum (seuil) pour les hits.

Les questions suivantes se posent:

1. Quelle est la distribution du nombre de hits pour une séquence requête de taille r, une séquence cible de taille c, une matrice de score M et un seuil S?

2. Quelle est la probabilité de trouver deux hits dans une bande de largeur b avec les mêmes hypothèses qu'à la question précédante?

Les méthodes utilisées feront intervenir de la combinatoire des mots et en particulier des propriétés d'autocorrélation de mots, ou de corrélations entre mots. Des résultats de dénombrement de mots dans des textes utilisent ces propriétés.