Utilisation de familles de graines pour la recherche de motifs par filtrage

Gregory Kucherov, Laurent Noé

laurent.noe@loria.fr

03.83.59.30.11

LORIA
615, rue du Jardin Botanique, BP 101,
54602 Villers-lès-Nancy Cedex

Résumé

Nous proposons une méthode de recherche de motifs de taille fixée m avec au plus k substitutions dans le texte. Cette méthode se base sur le *design* et l'utilisation de familles de graines espacées (nommées également Q-grams espacés). Burkhardt et Kärkkaïnen [1] avaient développé et analysé une approche basée sur les graines espacées en n'utilisant qu'une seule graine pour la recherche. Nous proposons ici un extension du filtre en réalisant le *design* simultané de plusieurs graines utilisées de manière disjonctive, permettant ainsi d'améliorer les taux de filtrage d'un facteur 20 sur des problèmes comme la recherche de fragments uniques à k erreurs près pour le design d'oligos.

Introduction

Les graines espacées (ou Q-grams espacés) ont été proposées en Bioinformatique dans l'algorithme implanté dans `FLASH` (Fast Look-Up Algorithm for String Homology [2]). Il s'agissait d'une méthode dérivée de celles proposées conjointement pour la reconnaisance de formes associée à des problèmes de vision. Le principe du filtrage avec perte (respectivement sans perte) consiste à éliminer de manière rapide des régions qui ont peu de chances (respectivement aucune chance) d'être similaires selon un critère fixé. Malgré son efficacité, le principe n'a pas été développé durant les années 90 pour les problèmes biologiques. Ce n'est que vers le début de la décennie que le concept est réapparu pour résoudre deux problèmes différents :

- la recherche exacte de fragments de taille m fixée ayant au plus k substitutions (que l'on notera par la suite problème (m, k)).
- la recherche heuristique de séquences redondantes, visant à améliorer la sensibilité des algorithmes d'alignement local de type heuristique (`FASTA`, `BLAST`, `Pattern-Hunter`,`YASS`,...).

Filtrage de texte

Le premier problème (recherche exacte par utilisation d'un filtre sans perte) a été étudié en détail par Burkhardt et Kärkkäinen. Il s'agissait de trouver une graine espacée permettant de résoudre des problèmes (m, k) déterminés, c'est à dire réaliser un filtre sans perte trouvant la totalité des instances de motifs ayant k erreurs de substitutions.

Un extension proposée ici pour cette méthode consiste à utiliser plusieurs graines (une famille de graines). Le but est d'augmenter la selectivité du filtre en utilisant des graines dites de poids élevé. (tout en se limitant à des tailles de familles raisonnables).

Le design des graines par la méthode proposée en [1] n'étant plus applicable en pratique dès que la taille de la famille atteint 3 graines ou plus, une approche heuristique a été adoptée (algorithme génétique en l'occurrence).

Résultats et Experiences

Sur un alphabet à 4 lettres, la méthode précédemment décrite donne des familles de graines ayant une meilleure sélectivité globale, comparée aux méthodes de filtrage les plus courantes [3]. On mesure la sélectivité δ comme la probabilité que deux positions aléatoires sur deux séquences i.i.d vérifient le critère du filtre.

TAB. 1 – familles de graines trouvées pour le problème $(m = 50, k = 5)$

w	$taille\ famille$	$exemple$	δ
12	1	111010011101001110 1 ou 1010100010000010101000100000101010001[1]	$5.96\ 10^{-8}$
14	2	11110101100111101011 10110011110101100111[2]	$7.45\ 10^{-9}$
15	3	10011010111111001101011 10111111001101011111 11110011010111111 0011	$2.79\ 10^{-9}$
16	4	1110110101110011111111 11010111001111111101101 11100111111011010111 11111110110101110011 1	$9.31\ 10^{-10}$
17	6	11010110011111110111101 10110011111110111101011 11111110111101011 00111 11101111010110011111111 11110101100111111110111 11001111111101111010110 01	$3.49\ 10^{-10}$

TAB. 2 – familles de graines trouvées pour le problème $(m = 32, k = 5)$

w	$taille\ famille$	δ
7	1	$6.10\ 10^{-5}$
8	2	$3.05\ 10^{-5}$
9	3	$1.14\ 10^{-5}$
10	4	$3.81\ 10^{-6}$
11	6	$1.43\ 10^{-6}$

Références

[1] S. Burkhardt, Juha. Kärkkäinen. Better Filtering with Gapped q-Grams, *Fundamenta Informaticae*, 23 :1001–1018, 2003.

[2] A. Califano, I. Rigoutsos. Flash : A fast look-up algorithm for string homology, *Proceedings of the 1st International Conference on Intelligent Systems for Molecular Biology*,56–64, 1993.

[3] P. Pevzner, M. Waterman. Multiple Filtration and Approximate Pattern Matching, *Algorithmica*,135–154, 1995.