Selection d'oligonucleotides spécifiques à l'aide de familles de graines

Gregory Kucherov, Laurent Noé

LORIA, Nancy

Résumé

Nous nous intéressons au problème de la sélection d'oligonucléotides spécifiques. Il s'agit d'évaluer sur une séquence cible S les fragments courts (de l'ordre de 20 à 50 bases) qui sont spécifiques à cette séquence : ces motifs doivent ne s'hybrider qu'avec S, et être suffisamment distants d'un jeu défini de séquences de fond B pour ne pas s'hybrider à l'une d'entre elles.

La méthode proposée pour la selection se base sur un filtrage sans perte. Ce dernier détermine tous les fragments de la séquence cible S qui n'ont aucune chance de s'hybrider avec une des séquences de fond B.

La méthode de filtrage proposée repose principalement sur une famille de graines espacées (motifs non contigus conservés) pour réaliser un index multiple.

Introduction

L'usage le plus courant des oligonucléotides (probes) est lié au puces à ADN où un oligonucléotide (court fragment d'ADN) est selectionné pour ne s'hybrider qu'avec un fragment d'EST spécifié. Les amorces de PCR (Polymerase Chain Reaction) peuvent également être considérées comme des oligonucléotides possédant la propriété d'initialiser la duplication de séquences à une position spécifiée.

Pour concevoir de tels oligonucléotides destinés à ne s'hybrider que sur une séquence spécifiée S, un nombre important de méthodes commencent par une étape de filtrage. Cette étape enumère les fragments de S et detecte ceux qui sont similaires avec un fragment de B, donc non adaptés pour être considérés comme oligonucléotides spécifiques.

Méthode

Différents algorithmes ont été proposés, basés sur des structures commes les arbres de suffixes [3, 2], les tables de suffixes [6, 5], et des critères de similarité comme le plus long facteur commun [8, 7].

Ici, nous nous intéressons plus généralement à un problème de pattern matching approché : deux fragments de taille fixée m sont considérés comme *similaires* si leur distance de Hamming est inférieure à un entier k donné. Un filtre sans

perte a été conçu pour detecter ce type de similarité(s), et ainsi garantir de trouver tous les fragments de S uniques à k substitutions près.

De manière à obtenir un algorithme efficace en pratique, une stratégie de filtrage basée sur une famille de graines espacées a été choisie. Il s'agit d'une extension de la méthode proposée par Burkhardt et Kärkkäinen [1]. L'extension à plusieurs graines permet des gains importants en sélectivité (facteur de l'ordre de 50 à 100).

Cependant, la recherche de la famille de graines optimales étant coûteuse pour un problème (m, k) donné, des méthodes heuristiques ont été mises en place de manière à obtenir des solutions satisfaisantes en temps raisonnable.

La première méthode se base sur un algorithme génétique pour la sélection des familles de graines. La seconde méthode utilise des solutions proposées dans [**?**] relatives à la résolution de problèmes (m', k) circulaires (avec $m' << m$).

Références

[1] Stefan Burkhardt and Juha Kärkkäinen. Better filtering with gapped q-grams. *Fundamenta Informaticae*, 56(1-2) :51–70, 2003. Preliminary version in Combinatorial Pattern Matching 2001.

[2] L. Kaderali and A. Schliep. Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics*, 18(10) :1340–1349, 2002.

[3] Stefan Kurtz, Enno Ohlebusch, Chris Schleiermacher, and Jens Stoye. Reputer : the manifold applications of repeat analysis. *Nucleic Acids Research*, 29(22) :4633–4642, 2001.

[4] A. Lefebvre, T. Lecroq, H. Dauchel, and J. Alexandre. FORRepeats : detects repeats on entire chromosomes and between genomes. *Bioinformatics*, 19(3) :319–326, 2003.

[5] S. Levy, L. Compagnoni, E.W. Myers, and G.D. Stormo. Xlandscape : the graphical display of word frequencies in sequences. *Bioinformatics*, 14(1) :74–80, 1998.

[6] F Li and GD. Stormo. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, 17 :1067–1076, 2001.

[7] Sven Rahmann. Fast and sensitive probe selection for DNA chips using jumps in matching statistics. In *IEEE Computer Society Bioinformatics Conference (CSB'03)*, pages 57–64.

[8] Sven Rahmann. Fast large scale oligonucleotide selection using the longest common factor approach. *Journal of Bioinformatics and Computational Biology*, 1(2) :343–361, 2003.