Sensibilité de graines espacées de type *subset seed*

Laurent Noé, Gregory Kucherov

LORIA, Nancy

Résumé

Les heuristiques d'alignement local à base de graines espacées (BLASTZ [6], PatternHunter [4, 3], YASS [5]) dépendent très fortement du choix des graines (expressions régulières particulières). L'évaluation de ces graines est faite à l'aide d'un algorithme de programmation dynamique coûteux. Nous considérons un modèle étendu de graines espacées (subset seeds), et nous proposons un algorithme efficace pour l'évaluation de la sensibilité de ces graines.

Introduction

Dans le cadre de la recherche heuristique d'alignements locaux, la qualité du résultat des méthodes dites à base de graines (expressions régulières particulières) est lié au choix de la graine. Une graine mal adaptée prise au hasard fait perdre en moyenne 30% des alignements significatifs trouvés par une graine optimale. De manière à établir une sélection des meilleures graines, nous sommes interessés par le calcul de leur sensibilité.

Les graines espacées classiques sont généralement évaluées à l'aide de méthodes de programmation dynamique sur un automate (Mandala [2, 7], PatternHunter [3]). Cependant, dès que le modèle est étendu (*subset seeds*), un coût supplémentaire apparaît, et le calcul devient alors vite rédhibitoire. Il est cependant possible d'obtenir un algorithme de programmation dynamique équivalent à celui des graines classiques, en choisissant une autre approche de construction de l'automate.

Subset seeds et automate induit

Lorsqu'un alphabet biologique est donné (nucléotides, acides aminés), la notion d'erreur (mismatch) entre lettres est relative : elle dépend des propriétés associées aux composés et à leurs éventuels substituts (par exemple on peut distinguer les "purines" et les "pyrimidines" parmi les 4 nucléotides, ou des propriétés comme "hydrophilie" et "hydrophobie" dans le cadre des acides aminés.

Ces propriétés sont prises en compte lors des processus d'évaluation des comparaisons (à l'aide de matrices de scores), mais sont rarement utilisés dans le processus même de recherche et d'indexation. Les subsets seeds reprennent cette idée de séparation en classes et étendent le modèle classique des graines espacées.

Par exemple, dans le cadre d'un alphabet nucléique, on peut séparer deux "familles" que sont les purines (`A` et `G`) et les pyrimidines (`T` et `C`). Ainsi une graine espacée composée sur un alphabet à deux lettres (distinction entre les matches obligatoires '#' et les jokers '_') peut être étendue par une lettre distinguant les purines et les pyrimidines. Ainsi la graine `#@#__##__#_##@#` formée sur cet alphabet détecte l'alignement suivant à la 17eme position : ses caractères (`#`) sont alignés avec des matches (`|`) et ses caractères `@`) correspondent soit à des matches (`|`), soit à des transitions (`.`) : alignement purine-purine ou pyrimidine-pyrimidine.

```
GGTGGCCAATTAACTGAGGCTGTTAGAAGACAGC
||||||::|.|:||.||.||.|||:|:|||||.:|
GGTGGCGTACTCACCGAAGCCGTTCGCAGGCGCC
                  #@#__##__#_##@#
```

Le design de telles graines s'avère plus difficile, de part l'explosion combinatoire rapide du nombre de motifs à considérer mais également du fait d'un calcul plus complexe pour chaque motif. Nous nous intéressons au deuxième problème et proposons une méthode plus efficace que celle couramment adoptée (Aho-Corasick [1, 2, 7]) pour générer l'automate des graines .

Références

[1] A. V. Aho and M. J. Corasick. Efficient string matching : An aid to bibliographic search. *Communications of the ACM*, 18(6) :333–340, 1975.

[2] J. Buhler, U. Keich, and Y. Sun. Designing seeds for similarity search in genomic DNA. In *Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB03), Berlin (Germany)*, pages 67–75. ACM Press, April 2003.

[3] Ming Li, Bin Ma, Derek Kisman, and John Tromp. PatternHunter II : Highly sensitive and fast homology search. *Journal of Bioinformatics and Computational Biology*, 2004.

[4] Bin Ma, John Tromp, and Ming Li. PatternHunter : Faster and more sensitive homology search. *Bioinformatics*, 18(3) :440–445, 2002.

[5] L. Noé and G. Kucherov. Improved hit criteria for DNA local alignment. *BMC Bioinformatics*, 5(149), october 2004.

[6] Scott Schwartz, James Kent, Arian Smit, Zheng Zhang, Robert Baertsch, Ross Hardison, David Haussler, and Webb Miller. Human–mouse alignments with BLASTZ. *Genome Research*, 13 :103–107, 2003.

[7] Yanni Sun and Jeremy Buhler. Designing multiple simultaneous seeds for DNA similarity search. In *Proceedings of the 8th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2004)*, pages 76–84. ACM Press, March 2004.