Graines espacées et recherche d'ARN non-codants

Arnaud Fontaine, Mathieu Giraud, Laurent Noé, Hélène Touzet

Bioinfo/Sequoia, LIFL, UMR CNRS 8022, USTL, INRIA Futurs, Villeneuve d'Ascq cedex
{fontaina, giraud, noe, touzet}@lifl.fr

Abstract: *Sequence comparison is widely used to help discovering novel non-coding RNAs in newly sequenced genomes. In this context, Blast-like homology search tools are of great interest. We show here that the usage of software based on "spaced seeds" has a positive impact on non-coding RNA identification.*

Keywords: non-coding RNAs, homology searches, spaced seeds, comparative genomics

1 Recherche d'ARN non-codants

Les ARN non-codants (ARNnc) jouent des rôles multiples et essentiels dans la cellule, participant à la synthèse protéique et à la régulation de l'expression des gènes [3]. Leur fonctionnalité dépend souvent davantage de leur organisation spatiale, conservée par l'évolution, que de leur séquence en nucléotides. Les meilleurs outils pour la détection d'ARNnc sont donc ceux qui tiennent compte de la structure [4]. Cependant, le recours à des outils de comparaison de séquences pour identifier des ARNnc par génomique comparative reste indispensable quand aucune information de structure n'est connue [2,8,10]. Traditionnellement, le logiciel Blast [1] et ses dérivés sont les plus utilisés. Est-il possible d'utiliser des méthodes plus sensibles afin de mieux explorer les génomes à la recherche d'ARNnc ?

2 Graines et recherche de régions conservées

Les programmes tels que Fasta ou Blast sont des heuristiques : ils recherchent tout d'abord des k-mers entièrement conservés, les *graines*, puis étendent ces graines pour construire des alignements. Le principe des graines a été amélioré avec l'introduction de *graines espacées* à la place des graines contiguës que sont les k-mers (Figure 1). Les graines espacées apportent une meilleure sensibilité, permettant ainsi de trouver des régions moins conservées, sans dégrader ni l'efficacité ni la spécificité [6].

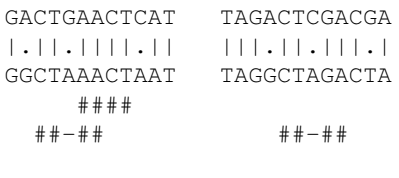
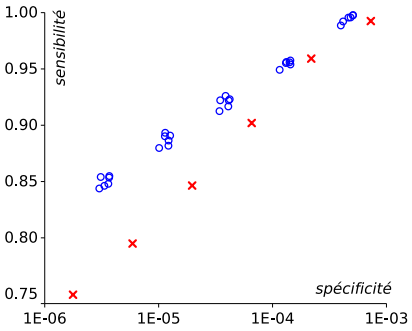


Fig. 1. Principe des graines espacées. Les deux alignements présentent une identité de $9/12 = 75\%$. On considère deux graines de poids 4 : la graine contiguë `####` et la graine espacée `##-##`. La graine contiguë ne détecte que le premier alignement, alors que la graine espacée détecte les deux. Le symbole `#` correspond à une position d'identité et le symbole `-` à une position quelconque.

Les ARNnc tendent à être moins conservés que les régions codantes, donc plus difficiles à déceler par génomique comparative. Le but de ce travail est d'évaluer le gain pratique que peuvent apporter les algorithmes à base de graines espacées. Pour cela, nous avons comparé systématiquement les résultats obtenus avec le logiciel Blast [1] et le logiciel Yass, qui implémente le principe des graines espacées [7]. Comme premier jeu de test, nous avons utilisé la base de données d'ARN non-codants RFAM [5], qui comprend 574 familles de petits ARN. Nous avons complété cette étude en reprenant les benchmarks proposés en 2007 par Freyhult *et al.*, composés de familles d'ARN de transfert, d'ARN ribosomiques 5S et d'ARN du spliceosome U5 [4]. Dans les deux cas, les expériences font apparaître de meilleurs résultats avec Yass qu'avec Blast (Figure 2).



	ARNt 1114 séquences	ARNnc U5 235 séquences	ARNr 5S 602 séquences
Blast	0.04	0.85	0.32
Yass	0.18	0.93	0.59

Fig. 2. À gauche, compromis spécificité-sensibilité entre les graines contiguës (Blast, croix) et les graines espacées (Yass, cercles) sur des alignements dans des familles de RFAM présentant plus de 50% de similarité. À droite, sensibilité de Yass et de Blast sur les benchmarks proposés par [4], avec une E-valeur de 10^{-4} et les arguments par défaut.

3 Annotation de génomes

Nous avons également étudié le gain des techniques à base de graines espacées à l'échelle de génomes bactériens en fonction de la distance évolutive entre les organismes. Comme point de départ, nous avons repris les travaux de [8] pour une recherche de candidats chez *E. coli* par comparaison avec quatre autres entérobactéries (*S. enterica*, *S. enteriditis*, *S. typhi* et *K. pneumoniae*). Le génome d'*E.coli* est utilisé comme génome pivot, et ses régions intergéniques sont alignées par paires avec les quatre autres génomes. Les zones sélectionnées sont celles qui sont alignées au moins $1,6$ fois plus que la moyenne sur *E.coli*. Avec une E-valeur théorique de 10^{-2}, les zones détectées par Blast contiennent ou intersectent 153 ARNnc, et celles détectées par Yass 155 (sur une référence de 160 ARNnc). Nous avons repris ce protocole en utilisant d'autres espèces plus éloignées (*L. pneumophilia*, *R. etli*, *G. sulfurreducens* et *M. tuberculosis*). Dans ce cas, le résultat de Yass reste stable, avec 153 ARNnc correctement identifiés, alors que Blast n'en repère plus que 109.

La méthode a été systématisée en un pipeline logiciel dédié à la découverte d'ARNnc (disponible sur demande) : comparaison deux à deux des régions intergéniques avec Yass, identification de régions conservées communes, et extraction de séquences candidates d'intérêt avec le logiciel d'inférence de structure secondaire caRNAc [9].

References

[1] S. F. Altschul and al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res., 25:3389-3402, 1997

[2] G. Bejerano, D. Haussler, M. Blanchette, *Into the heart of darkness: large-scale clustering of human non-coding DNA*, Bioinformatics, 20 Suppl 1:I40-I48, 2004

[3] S. R. Eddy, *Non-coding RNA genes and the modern RNA world*, Nat Rev Genet., 2(12):919–29, 2001

[4] E. Freyhult, J. Bollback, P. Gardner, *Exploring genomic dark matter: a critical assessment of the performance of homology search methods on non-coding RNA*, Genome Research, 17:117-125, 2007

[5] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, A. Bateman, *Rfam: annotating non-coding RNAs in complete genomes*, Nucleic Acids Res., 33:D121–D124, 2005

[6] B. Ma, J. Tromp, M. Li, *PatternHunter: faster and more sensitive homology search*, Bioinformatics 18(3):440-5, 2002

[7] L. Noé, G. Kucherov, *YASS: enhancing the sensitivity of DNA similarity search*, Nucleic Acids Res., 33:W540-W543, 2005

[8] E. Rivas, R. J. Klein, T. A. Jones, S. R. Eddy, *Computational identification of noncoding RNAs in E. coli by comparative genomics*, Curr Biol. 11(17):1369-73, 2001

[9] H. Touzet, O. Perriquet, *CARNAC: folding families of related non coding RNAs*, Nucleic Acids Res. 142, 2004

[10] S. Washietl, I. L. Hofacker, M. Lukasser, A. Hüttenhofer, P. F. Stadler, *Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome*, Nat Biotechnol. 23(11):1383-90, 2005