Automate des graines sous-ensemble

Gregory Kucherov[1], Laurent Noé[1], Mikhail Roytberg[2]

1 INRIA Futurs, 40 avenue Halley, 59650 Villeneuve d'Ascq, France,
{gregory.kucherov,laurent.noe}@lifl.fr

2 Institute of Mathematical Problems in Biology, Pushchino, Moscow Region, 142290, Russia,
mroytberg@mail.ru

Dans le cadre de la recherche de similarités dans les séquences d'ADN, le modèle des *graines espacées* [1, 2] et ses généralisations comme par exemple les *graines sous-ensemble* [3] ont apporté un gain en sensibilité significatif, comparé au modèle traditionnel des *graines contiguës* [4]. Une graine espacée prend en compte, à une position donnée sur une séquence, non pas la totalité des lettres sur un intervalle, mais seulement un sous-ensemble fixé par une liste d'indices (figure 1). Si les deux séquences d'une similarité possèdent au moins une graine commune, alors cette similarité est détectée par l'heuristique.

Figure 1: Fragment contigu et espacé

De manière intuitive, l'approche par graines espacées autorise donc certaines mutations (les substitutions) à apparaître sous les positions non considérées par la liste d'indices.

Cependant le choix du *motif de graine* est désormais devenu prépondérant. Il possède d'une part un degré de liberté bien plus élevé. D'autre part, l'évaluation de la sensibilité (associée à chacun des motifs), auparavant polynomiale, est dorénavant exponentielle en la taille de la graine pour chaque motif [5, 6]. L'algorithme couramment employé pour réaliser cette évaluation se résume à un calcul de probabilités, par programmation dynamique, sur un automate du langage reconnu par une graine donnée [7]. La taille de l'automate influence grandement la complexité finale, et le choix de minimiser cet automate, ou de le construire suffisamment petit, se pose...

Nous avons étudié l'automate des *graines sous-ensemble* proposé dans [3] : nous présentons dans cet exposé ses propriétés, en particulier celles relatives à sa taille, ainsi qu'une implémentation efficace pour le construire en temps linéaire. Nous présentons également des expériences réalisées sur cet automate, et montrons enfin qu'il peut être utilisé dans un cadre plus général que celui de l'indexation par graines.

References

[1] Ma, B., Tromp, J., Li, M.: PatternHunter: Faster and more sensitive homology search. Bioinformatics **18**(3) (2002) 440–445

[2] Burkhardt, S., Kärkkäinen, J.: Better filtering with gapped q-grams. In: Proceedings of the 12th Symposium on Combinatorial Pattern Matching (CPM). Volume 2089 of Lecture Notes in Computer Science. (2001) 73–85

[3] Kucherov, G., Noé, L., Roytberg, M.: A unifying framework for seed sensitivity and its application to subset seeds. JBCB **4**(2) (2006) 553–569

[4] Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D.: Basic Local Alignment Search Tool. Journal of Molecular Biology **215** (1990) 403–410

[5] Li, M., Ma, B., Kisman, D., Tromp, J.: PatternHunter II: Highly sensitive and fast homology search. Journal of Bioinformatics and Computational Biology **2**(3) (2004) 417–439

[6] Nicolas, F., Rivals, E.: Hardness of optimal spaced seed design. In: Proceedings of the 16th Annual Symposium on Combinatorial Pattern Matching (CPM), Jeju Island (Korea). Volume 3537 of LNCS. (2005) 144–155

[7] Buhler, J., Keich, U., Sun, Y.: Designing seeds for similarity search in genomic DNA. In: Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB). (2003) 67–75