Alignement de séquences avec des opérations non-commutatives

Eric Rivals, Sèverine Bérard
LIRMM, 161 rue Ada,
34392 Montpellier Cedex 5
FRANCE
rivals@lirmm.fr

L'étude des génomes a révélé l'abondance des séquences répétées, notamment dans le cas des organismes complexes : elles représentent environ 40% du génome humain [1]. La cause de leur apparition ainsi que leur évolution sont encore mal comprises. Les séquences constituées de segments répétés adjacents le long du chromosome, séquences dites "répétées en tandem", sont sujettes à un mécanisme particulier d'évolution. Elles subissent des duplications ou des pertes en tandem du motif répété, ce qui fait varier leur longueur. Les mutations ponctuelles (insertion, délétion et substitution d'un symbole) altèrent les copies du motif. La combinaison des deux types d'événements produit une suite de motifs adjacents, légèrement différents les uns des autres. Ainsi, peut-on observer chez deux individus d'une même espèce des suites différentes de motifs. Nous considérons le problème d'aligner optimalement deux suites de motifs pour mettre en évidence la série d'événements qui transforment l'une en l'autre.

Dans le modèle classique d'alignement de séquences [3], seules les mutations ponctuelles sont considérées. Dans notre contexte, les motifs sont les symboles de la suite et nous autorisons les mutations ponctuelles sur les motifs ainsi que les événements de duplication et de perte en tandem d'un motif. À chaque opération est associée un coût fixé et le coût de l'alignement est la somme des coûts des opérations qu'il contient. Contrairement au cas classique, nos opérations ne sont plus commutatives ; il s'ensuit qu'aligner des paires de préfixes de plus en plus longs des deux suites par programmation dynamique (comme pour l'alignement classique [3]) ne suffit plus à calculer l'alignement optimal. Nous proposons un algorithme exact qui combine programmation dynamique et algorithmique de graphes avec une complexité

cubique en fonction de la longueur des suites (cf. [2]).

En conclusion, nous discuterons une application sur des séquences génétiques et des perspectives algorithmiques telles que la prise en compte de coûts variables, de duplication de plusieurs motifs, ou la comparaison multiple.

Références

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson. *Molecular Biology of the Cell*. Garland, New York, 1983.

[2] Sèverine Bérard and Eric Rivals. Comparison of minisatellites. *J. of Computational Biology*, 10(3-4) :357–372, 2003.

[3] David Sankoff and Joseph B. Kruskal, editors. *Time Warps, String Edits and Macromolecules : the Theory and Practice of Sequence Comparison*. CSLI Publications, second edition, 1999.