Comparaison de minisatellites

Sèverine Bérard et Éric Rivals
L.I.R.M.M., UMR CNRS 5506
161, rue Ada, 34392 Montpellier Cedex 5
email: {berard, rivals}@lirmm.fr

Nous présentons ici une méthode nouvelle pour comparer des cartes de minisatellites.
Les minisatellites (ms) appartiennent à la classe des séquences répétées en tandem. Ces séquences sont constituées de répétitions en tandem d'un motif unitaire. Selon la taille du motif et de la répétition, on distingue les satellites, les minisatellites et les microsatellites. Comme les autres séquences répétées en tandem, les minisatellites sont polymorphes. En effet, ces séquences subissent des fluctuations de leur nombre de motifs. Ce phénomène est dû à la *duplication en tandem*, un événement qui ajoute un motif copié à côté de l'original, et de l'événement inverse, la *délétion en tandem*. Comme pour les microsatellites, la variation de longueur des minisatellites est impliquée dans plusieurs maladies comme le diabète, l'épilepsie et le cancer. Du fait de de leur polymorphisme, les minisatellites sont utilisés pour la cartographie génétique et les études médico-légales.

Individu 1

Evt	Séquence
	a a a a a
M	a **e** a a a

Individu 2

Evt	Séquence
	a a a a a
M	a a a **b** a
2*A	a a a b **b** **b** a
I	a a a b b **c** b a

(a) Exemple d'évolution d'un minisatellite chez 2 individus. La première colonne représente l'événement, la seconde le résultat de cet événement sur la séquence.

<table>
<tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td></tr>
<tr><td>I_1</td><td>a</td><td>e</td><td>a</td><td>a</td><td>–</td><td>–</td><td>–</td><td>a</td></tr>
<tr><td></td><td>|</td><td>]</td><td>|</td><td>]</td><td>\</td><td></td><td>\</td><td>|</td></tr>
<tr><td>I_2</td><td>a</td><td>a</td><td>a</td><td>b</td><td>b</td><td>c</td><td>b</td><td>a</td></tr>
</table>

(b) Alignement de I1 et I2.

FIG. 1 – Dans cette figure nous considérons deux cartes de minisatellites, I_1=aeaaa et I_2=aaabbcba. (a) détaille l'évolution du ms chez deux individus. (b) montre comment nous souhaiterions les aligner. Dans (b), '|', ']' et '\' représentent respectivement une identité, une mutation et une amplification ; la mise en correspondance de '-' et d'un symbole représente une insertion.

En 1991, Jeffreys et ses collaborateurs ont mis au point une technique PCR spécifique aux minisatellites, la MVR-PCR [2], où MVR signifie Minisatellites Variant Repeat. Cette méthode fournit une *carte du minisatellite*, càd la séquence de ses motifs, où chaque motif est représenté par un symbole. Pour exploiter pleinement les informations évolutives contenues dans ces cartes, il est crucial de pouvoir les comparer de manière automatique. Nous proposons un algorithme d'alignement qui prend en compte les événements de duplication et de délétion en tandem. Nous nous plaçons dans le cadre d'un modèle évolutif symétrique et unaire.
Notre modèle considère cinq événements évolutifs s'appliquant aux motifs, la mutation (M), l'insertion (I), la suppression (S), l'amplification (A) et la contraction (C). Mutation, insertion et suppression sont les événements que l'on considère traditionnellement dans l'alignement de séquences mis à part qu'ici ils s'appliquent à des motifs. Les deux événements spécifiques sont l'amplification et la contraction, ils correspondent respectivement à la duplication et à la délétion en tandem. Par exemple, la séquence *abc* subit l'amplification du motif b, puis sa contraction :

$$\textbf{amplification} : abc \longrightarrow abbc \qquad\qquad \textbf{contraction} : abbc \longrightarrow abc.$$

Notons que l'amplification insère un motif x à une position $i+1$ seulement si un motif x est déjà dans la séquence à la position i, et que la contraction supprime un motif y à la position j seulement si un autre motif y se trouve à la position $j+1$. Notre modèle est unaire car l'amplification (resp. la contraction) ajoute (resp. retire) un seul motif à la fois. Pour compléter ce modèle, nous avons besoin d'un critère quantitatif permettant d'estimer la similarité des cartes. Aussi, à chaque opération est associée un coût réel positif. Une séquence d'événements qui transforme une carte s en une carte r est appelé un *alignement*, un exemple est donné à la Figure 1(b). Le coût d'un alignement est la somme des coûts des opérations qui le compose. La fréquence plus élevée des amplifications et des contractions par rapport aux autres opérations se traduit par le fait qu'elles ont un coût plus faible. Cependant, ces deux opérations sont soumises à conditions. La recherche de l'alignement optimal est difficile en raison de la non-commutativité des opérations.

L'exemple de la Figure 1 souligne l'importance de l'ordre des opérations. Dans 1(a) nous montrons l'évolution d'un ms chez deux individus, conduisant à deux cartes, I_1 et I_2. Il y a deux manières d'aligner ces cartes, et notamment la position 4 de I_1 avec les positions 4 à 7 de I_2.

1. Muter le a, à la position 4 de I_1 en b, amplifier le b ainsi obtenu, insérer un c en position 6, puis insérer un b en position 7.
2. Muter le a, à la position 4 de I_1 en b, amplifier 2 fois le b ainsi obtenu, puis insérer le c en position 6. (cas représenté dans 1(b))

Notons que le b de la position 7 dans I_2 provient d'une amplification dans le cas 2 et d'une insertion dans le cas 1, or l'amplification est moins coûteuse que l'insertion. Le reste des opérations étant identique, la deuxième manière est donc moins coûteuse que la première. Pour obtenir le meilleur alignement, il faut faire les amplifications avant l'insertion du c, cela montre que les opérations ne sont pas commutatives, par conséquent leur ordre d'application est important. Nous appelons *arche* ces sous-séquences où l'ordre d'application des opérations pour obtenir un coût optimal n'est pas un ordre "de gauche à droite", et nous les alignons avec un seul motif de la séquence en face. Par exemple, la sous-séquence de I_2 commençant à la position 4 et terminant à la position 7, est une arche de I_2. Nous pouvons traiter les arches de manière indépendante du reste de l'alignement. La création du concept d'arches implique des définitions et des méthodes (pouvant se traduire en termes de graphes de recouvrement) que nous ne pouvons détailler ici.

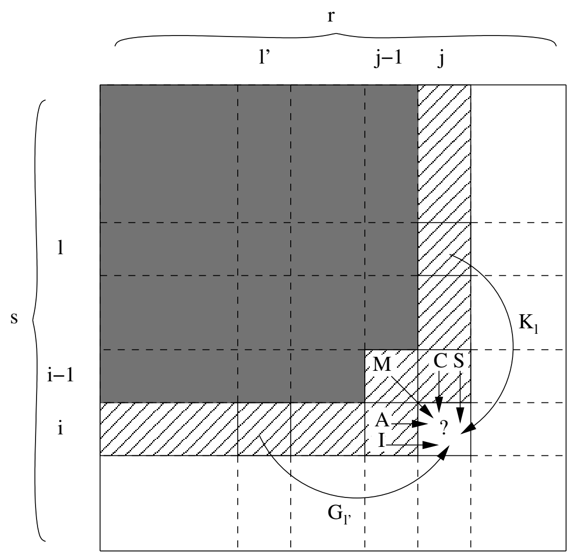

FIG. 2 – Dépendances dans la matrice de programmation dynamique.

Notre algorithme est un algorithme de programmation dynamique. Il diffère des algorithmes de programmation dynamique classiques dans les alignements de séquences car il considère des opérations supplémentaires, l'amplification, la contraction et les opérations d'arches : génération d'arches (G) et compression d'arches (K). Les dépendances sont illustrées dans la Figure 2. La complexité en temps de l'algorithme est en $O(n^4)$, où n est la taille de la plus longue des deux séquences à aligner.

Nous avons appliqué notre algorithme au minisatellite humain MSY1 qui est situé sur le chromosome Y [3]. Le jeu de donnée contenait 609 cartes d'individus répartis en 27 haplogroupes. Nous utilisons la matrice des distances entre individus pour reconstruire des arbres évolutifs pour l'ensemble des individus ou pour chacun des haplogroupes. Les résultats sont en cours d'investigation.

Pour plus de détails vous pouvez vous référer à [1].

Références

[1] S. Bérard and E. Rivals. Comparison of Minisatellites. In *Proc. of the 6th RECOMB*, Washington, USA, 2002. ACM Press.

[2] A. J. Jeffreys, A. MacLeod, K. Tamaki, D. L. Neil, and D. G. Monckton. Minisatellite repeat coding as a digital approach to DNA typing. *Nature*, 354(6350) :204–9, 1991.

[3] M. A. Jobling, N. Bouzekri, and P. G. Taylor. Hypervariable digital DNA codes for human paternal lineages : MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). *Hum Mol Genet*, 7(4) :643–53, 1998.