

ASSIRC

Accelerated Search for Similarity Regions in Chromosomes

P. Vincens, L. Buffat, C. André, J.-P. Chevrolat,
J.-F. Boisvieux et S. Hazout

Bioinformatics **14**(8) (1998) 715-725.

3 étapes :

1. localisation de courts fragments identiques dans les 2 séquences par hachage ;
2. extension des fragments par une marche aléatoire ;
3. alignement des régions sélectionnées en utilisant BESTFIT.

1. Hachage

Les facteurs de longueur k sont codés par la fonction

$$\text{code}(w[0..k-1]) = \sum_{i=0}^k 4^{k-i} \text{rang}(w[i])$$

avec $\text{rang}(\text{A}) = 0$, $\text{rang}(\text{T}) = 1$, $\text{rang}(\text{G}) = 2$ et $\text{rang}(\text{C}) = 3$

On utilise 2 tables à 4^k éléments

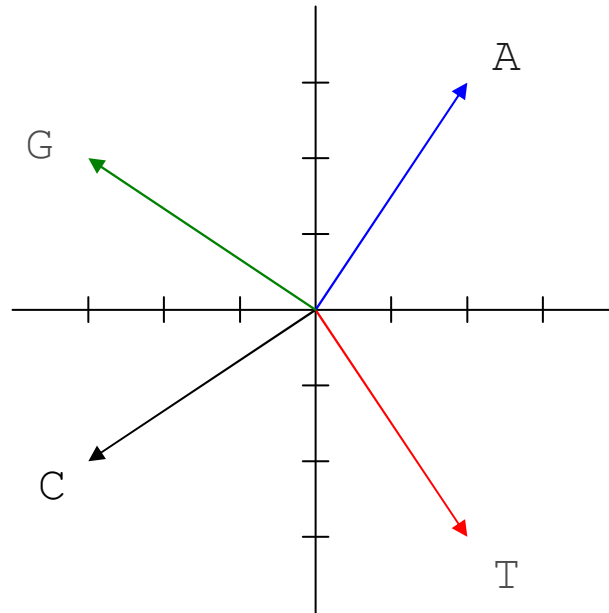
$$position_x[c] = \{ i \mid code(x[i..i+k-1]) = c \}$$

$$position_y[c] = \{ j \mid code(y[j..j+k-1]) = c \}$$

$$couple[c] = \{ (i,j) \mid i \in position_x[c] \text{ et } j \in position_y[c] \}$$

2. Extension

Transformation d'une séquence en courbe :
chaque nucléotide entraîne un déplacement sur le plan



Transformation injective

$$\begin{aligned} \text{abscisse}(u[0..i]) &= 2(|u[0..i]|_A + |u[0..i]|_T) \\ &\quad - 3(|u[0..i]|_C + |u[0..i]|_G) \end{aligned}$$

$$\begin{aligned} \text{ordonnée}(u[0..i]) &= 3(|u[0..i]|_A - |u[0..i]|_T) \\ &\quad - 2(|u[0..i]|_C - |u[0..i]|_G) \end{aligned}$$

Pour mesurer la similarité entre 2 séquences u et v de même longueur, on calcule leur distance euclidienne après transformation.

Les couples $couple[c]$ sont étendus à droite et à gauche jusqu'à ce que τ distances consécutives dépassent un seuil D .

Lorsque les régions de similarité atteignent une longueur L_{min} , elles sont stockées dans une base de données. Si il y a chevauchement avec une région déjà présente alors celle-ci est remplacée par la concaténation des 2 régions.

3. Alignement

Alignement des régions similaires.

Validation expérimentale

- plus k augmente plus le nombre de régions détectées est faible, et donc plus les calculs sont rapides ;
- plus D augmente, plus le temps d'exécution est grand.