

Répétitions en tandem

Définitions

Un mot $w = w_0w_1\dots w_{k-1}$, avec $k > 1$, est :

- une répétition en tandem exacte si $w_i = w_j$ pour $0 \leq i, j \leq k-1$;
- une répétition en tandem approchée avec motif consensus s'il existe $x \in A^*$ tel que $score(w_i, x) \geq seuil$;
- une répétition en tandem approchée deux à deux si $score(w_i, w_j) \geq seuil$ pour $0 \leq i, j \leq k-1$ et $i \neq j$;
- une répétition en tandem approchée avec évolution si $score(w_i, w_{i+1}) \geq seuil$ pour $0 \leq i < k-1$.

Les w_i sont appelés des motifs.

Référence

Tandem Repeats Finder

Gary Benson

Nucleic Acids Research 27(2) (1999) 573-580

Calcul des répétitions en tandem (RT) dans une séquence y de longueur n .

2 paramètres :

- p_M : probabilité d'égalité ;
- p_I : probabilité d'insertion/délétion.

2 phases :

- détection ;
- analyse.

Phase de détection

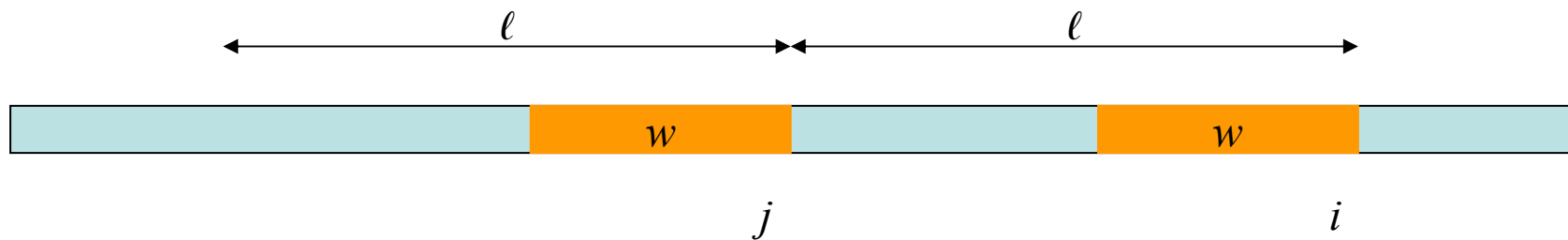
Pour un entier k , à l'aide d'une fenêtre glissante on détermine la liste des positions de chaque mot de longueur k dans la séquence y :

$$H_w = \{ j \mid w = y[j..j+k-1] \} \text{ pour } w \in A^k$$

Il y a 4^k mots de longueur k sur l'alphabet

$$A = \{ A, C, G, T \}.$$

Quand une position i est insérée dans H_w , les longueurs $\ell = i - j$, avec $j \in H_w$, sont des longueurs possibles de motifs de RT.



La position i est alors insérée dans une liste D_ℓ et toutes les positions inférieures à $j = i - \ell$ en sont supprimées.

Les listes $D_{\ell \pm \Delta\ell}$ sont également mises à jour de la même façon.

$\Delta\ell$ est déterminé en étudiant une marche aléatoire en une dimension ($\Delta\ell = \lfloor (p_1 \times \ell)^{1/2,3} \rfloor$).

Cela permet de prendre en compte les insertions et les délétions.

Les motifs candidats de longueur ℓ sont soumis à deux tests statistiques :

- la somme des faces ;
- la taille apparente.

S'ils passent ces deux tests alors on passe à la phase d'analyse.

Phase d'analyse

Dans la phase d'analyse un motif candidat de longueur ℓ est aligné avec son voisinage dans la séquence y en utilisant la programmation dynamique « en boucle » (*wraparound dynamic programming*).

Critère de la somme des faces

Les deux motifs testés doivent comporter suffisamment d'identité.

On utilise un modèle de Bernouilli.

Lorsqu'on aligne deux mots de même longueur on associe une égalité à face (F) et une inégalité à pile (P).

Exemple

A	T	G	C	A	T
A	T	C	G	A	T
<hr/>					
F	F	P	P	F	F

Critère de la somme des faces

Soit R_{ℓ,k,p_M} le nombre total de faces dans les suites consécutives de k faces ou plus dans une séquence iid (indépendante et identiquement distribuée) de Bernoulli de longueur ℓ avec une probabilité de face égale à p_M .

La distribution de R_{ℓ,k,p_M} est bien approximée par une loi normale.

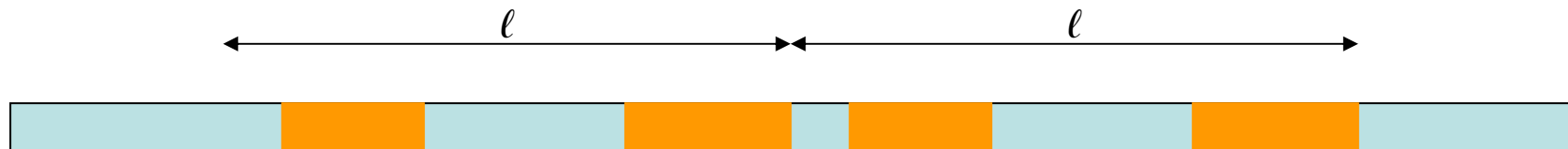
Pour ce critère on utilise la loi normale pour déterminer le plus grand nombre h tel que dans 95% des cas

$$R_{\ell,k,p_M} \geq h.$$

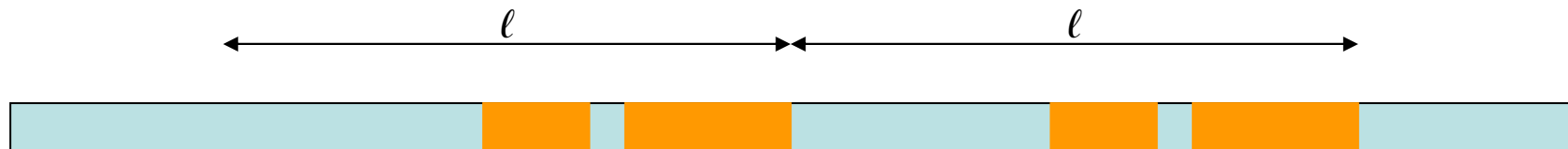
Les deux motifs candidats doivent comporter au moins h égalités intervenant uniquement dans des fenêtres de longueur k .

Critère de la taille apparente

Ce critère permet de distinguer les RT des répétitions dispersées.



RT



répétition dispersée

Critère de la taille apparente

Soit S_{ℓ,k,p_M} la longueur entre le premier et le dernier facteur de k faces dans une séquence iid de Bernoulli de longueur ℓ avec une probabilité de face égale à p_M .

La distribution de S_{ℓ,k,p_M} est déterminée par simulation car les motifs doivent d'abord satisfaire le critère de la somme des faces.

Critère du temps d'attente

Ce critère est utilisé pour déterminer la valeur de k : si k est grand cela diminue les chances de trouver des répétitions, si k est petit cela augmente le temps d'exécution. Ce critère permet de trouver un équilibre entre le temps d'exécution et la sensibilité.

Pour cela on étudie la distribution de la variable T_{k,p_M} égale au nombre de tirage iid de Bernouilli avec une probabilité de face égale à p_M jusqu'à obtenir k faces consécutifs.