Des arbres de suffixes aux vecteurs de suffixes

Élise Prieur et Thierry Lecroq

UNIVERSITÉ DE ROUEN

A.B.I.S.S.

Introduction

Les vecteurs de suffixes ont été introduit par Monostori [1, 2, 3]. C'est une structure de données alternatives aux arbres de suffixes. On considère un vecteur de suffixes comme une suite de boîtes placées à certaines positions sur le mot y.

Notations :

- $\mathcal{A}_C(\mathit{Suff}(y))$: arbre de suffixes d'un mot y de longueur n sur l'alphabet A ;
- $(\mathit{position}, \mathit{longueur})$: représentation des étiquettes des branches de $\mathcal{A}_C(\mathit{Suff}(y))$;
- $\delta(p, (i, \ell)) = q$: transition du nœud p vers le nœud q étiquetée (i, ℓ) ;
- $\textsc{Cible}(p, a)$: état q tel qu'il existe une branche de p vers q dont l'étiquette commence par a ;
- $s(au) = u$: lien suffixe du nœud au est le nœud u ;
- $\mathit{posd}(u, y)$: position droite de la première occurrence de u dans y ;
- $\mathcal{V}(y)$: vecteur de suffixes du mot y ;
- B_j : boîte de $\mathcal{V}(y)$ située à la position j de y.

De l'arbre au vecteur

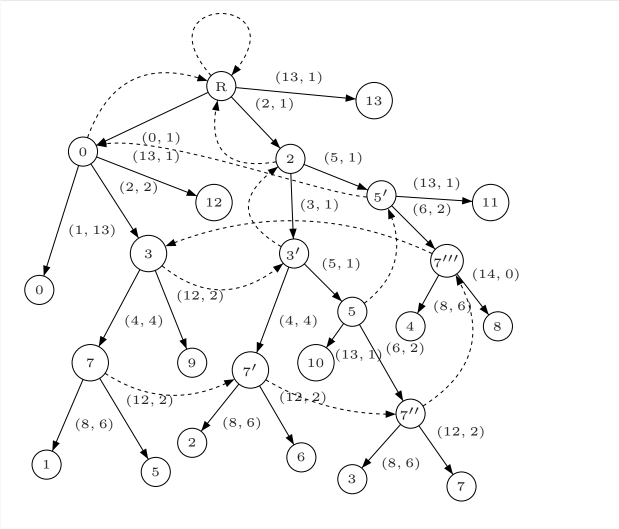
Soit $u \in \mathcal{A}_C(\mathit{Suff}(y))$, $j = \mathit{posd}(u, y)$ alors il existe B_j avec une ligne h telle que :

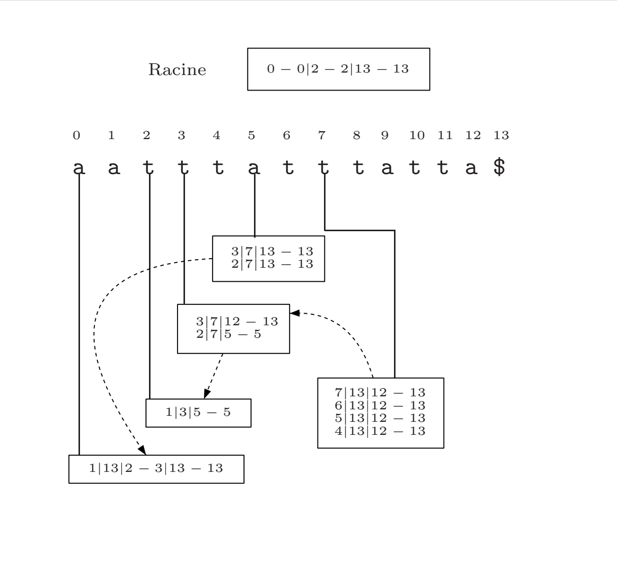
- $B_j[h, 0] = |u|$;
- $B_j[h, 1] = \mathit{posd}(\textsc{Cible}(u, y[j+1]), y)$;
- $\forall a \in A \setminus \{y[j+1]\}$ telle que $\textsc{Cible}(u, a)$ est définie, on a une transition $L[i] \in B_j[h, 2]$ telle que :
  - $L[i].f = \mathit{posd}(\textsc{Cible}(u, a), y)$ et
  - $L[i].d = L[i].f - |\textsc{Cible}(u, a)| + 1 + |u|$.

Exemple :
Pour $u = \texttt{att}$ soit le nœud 3 de l'arbre, on a *posd*(u,y)=3. D'où, pour $h = 0$:

- $B_3[0, 0] = |\texttt{att}| = 3$
- $B_3[0, 1] = \mathit{posd}(\textsc{Cible}(u, y[4]), y) = \mathit{posd}(\texttt{atttatt}, y) = 7$
- u possède une transition $(12, 2)$ donc
  - $|L| = 1$
  - $L[0].f = \mathit{posd}(\textsc{Cible}(u, a), y) = \mathit{posd}(\texttt{atta\$}, y) = 13$
  - $L[0].d = L[0].f - |\textsc{Cible}(u, a)| + 1 + |u| = 13 - 5 + 1 + 3 = 12$

Exemple pour le mot $y = \texttt{aatttatttatta\$}$





Du Vecteur à l'arbre

Soit B_j la boîte à la position j avec une ligne h telle que $B_j[h, 0] = \mathit{prof}$, $B_j[h, 1] = \mathit{cn}$ et $B_j[h, 2] = L$. Il existe un nœud u représentant $y[j - \mathit{prof} + 1..j]$ dans $\mathcal{A}_C(\mathit{Suff}(y))$ tel que :

- $\mathit{cn} = \mathit{posd}(\textsc{Cible}(u, y[j+1]), y)$;
- $\forall L[i] \in L$ telle que $L[i] = (d, f)$, il existe une branche $\delta(u, (L[i].d, L[i].f - L[i].d + 1)) = uy[L[i].d..L[i].f]$.

De plus, $s(y[j - B_j[h, 0] + 1..j]) = y[j - B_j[h+1, 0] + 1..j]$ pour h et $h+1$ deux lignes consécutives de B_j.

Exemple :
La ligne 0 de la boîte en position 5 représente le facteur $y[5-3+1..5] = \texttt{tta}$. Celui-ci équivaut au nœud 5 de $\mathcal{A}_C(\mathit{Suff}(y))$. Les informations de cette ligne du vecteur nous donne l'existence de :

- la branche de $\texttt{tta}$ vers un nœud $p = \texttt{tta}y[5..7] = \texttt{ttatt}$ avec $\mathit{posd}(p, y) = 7$ car $B_5[0, 1] = 7$;
- la branche étiquetée $(13, 1)$ allant vers une feuille car $B_5[0, 2] = (13, 13)$.

Le lien suffixe de cette ligne est la ligne 1, correspondant au nœud $5'$ de $\mathcal{A}_C(\mathit{Suff}(y))$.

Comptage

Le nombre d'occurrences d'un facteur u de y reconnu par une ligne du vecteur est $\mathit{nbOcc}(u) = \sum_L \mathit{nf}(L[i]) + \mathit{nf}((j+1, \mathit{cn}))$, où $\mathit{nf}(t)$ est le nombre de feuilles provenant de la transition $t = (t.d, t.f)$.

$$\mathit{nf}(t) = \begin{cases} 1 & \text{si } t.f = n; \\ \mathit{nbOcc}(v) & \text{sinon (avec } v \text{ tel que } \mathit{posd}(v, y) = t.f). \end{cases}$$

On peut en déduire le nombre de répétitions des facteurs d'une longueur donnée.

Exemple :

Boîte en position 7, 3ème ligne (même calcul pour les autres lignes) :

$\mathit{nbOcc}(\texttt{ttatt}) = \mathit{nf}((8, 13)) + \mathit{nf}((12, 13)) = 2$ car $|y| = 13$.

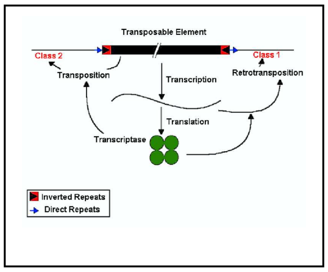
Boîte en position 3, 1ère ligne :

$\mathit{nbOcc}(\texttt{tta}) = \mathit{nf}((6, 7)) + \mathit{nf}((13, 13)) == \mathit{nbOcc}(\texttt{ttatt}) + 1 = 3$.

Perspectives

Utilisation de l'outil de comptage pour la détection de répétitions exactes dans les séquences biologiques afin d'aider à la détection d'éléments transposables.

Éléments transposables : fragments d'ADN capables de se déplacer d'un endroit du génome à un autre. On voit sur la figure ci-contre qu'ils sont entourés d'éléments répétés. On distinguera les rétrotransposons qui sont de la famille des rétrovirus et peuvent donc s'autodupliquer.



Références

[1] K. Monostori, A. B. Zaslavsky, I. Vajk. Suffix Vector : A Space-Efficient Suffix Tree Representation. In Proceedings of the 12th International Symposium on Algorithms and Computation, P. Eades and T. Takaoka eds., volume 2223 of Lecture Notes in Computer Science, pages 707-718, Christchurch, New Zealand, 2001. Springer Verlag.

[2] K. Monostori, A. B. Zaslavsky, H. W. Schmidt. Suffix Vector : Space- and Time-Efficient Alternative to Suffix Trees. In CRPITS '02 : Proceedings of the 25th Australasian Computer Science Conference, volume 4, pages 157-165, Melbourne, 2002. Australian Computer Society, Inc.

[3] K. Monostori. Efficient Computational Approach to Identifying Overlapping Documents in Large Digital Collections. PhD Thesis, Monash University, 2002.