**DM 5 – A rendre en TD la semaine du 10 janvier**

**KMP**

On s'intéresse à la recherche d'un motif $m$ dans un texte $t$. On notera $\ell$ la longueur de $m$, et $n_t$ celle de $t$. Etant donné un mot $w$, on notera $w[i..j]$ le sous-mot $w_i \dots w_j$ (par convention, $w[i..j] = \varepsilon$ si $i > j$).

**1.** Donner un algorithme naïf résolvant ce problème. Quelle est sa complexité ?

Soit $A = \langle Q, q_0, F, \Sigma, \delta \rangle$ un automate fini déterministe, et $\phi$ sa fonction d'état final ($\phi(w) \in Q$ est l'état dans lequel on arrive en partant de l'état initial et en lisant le mot $w$). On dit que $A$ est un automate de recherche de $m$ lorsque $Q = \{0, 1, \dots, \ell\}$ et

$$\phi(w) = \max\{q : m[1..q] \text{ est suffixe de } w\}.$$

**2.** Donner l'automate de recherche du motif *abac*.

**3.** En supposant qu'on dispose d'un automate de recherche de $m$, donner un algorithme de recherche de $m$ dans un texte. Justifier sa correction, et donner sa complexité.

**4.** Soit $\sigma(w) = \max\{k : m[1..k] \text{ est un suffixe de } w\}$. Montrer que $\delta(q, a) = \sigma(m[1..q] \cdot a)$.

**5.** En déduire un algorithme de calcul d'un automate de recherche de $m$ en $O(\ell^3)$

Nous allons maintenant essayer d'améliorer cette complexité. Un *bord* d'un mot $w$ est un sous-mot strict de $w$ qui soit à la fois son préfixe et son suffixe. On définit $\pi(q)$, pour $q \leq \ell$, comme étant la taille du plus grand bord de $m[1..q]$.

**6.** Que vaut $\delta(q, a)$ si $m_{q+1} = a$ ? Si $q = m$ ou $m_{q+1} \neq a$, montrer que $\delta(q, a) = \delta(\pi(q), a)$.
En supposant la fonction $\pi$ connue, en déduire un algorithme linéaire de calcul de l'automate de recherche.

**7.** Calculer les valeurs de $\pi$ pour $m = abcababcabcabb$.

**8.** Montrer que $\pi(q) > 0 \implies \exists i,\ \pi(q) = \pi^i(q-1) + 1$.
En déduire une définition inductive de $\pi(q)$.

**9.** Trouver un algorithme linéaire calculant la fonction de suffixe.

**10.** Vous venez de re-découvrir l'algorithme de *Knuth-Morris-Pratt*. Quelle est sa complexité ?