Motifs et sous-chaînes

Épreuve pratique d'algorithmique et de programmation

Juillet 2002

Le sujet étudie des algorithmes de recherche de motifs dans une chaîne de caractères. Les questions sont (très librement !) inspirés de problèmes rencontrés en génomique.

On s'intéresse à une longue chaîne C de caractères représentés par les nombres 0, 1, 2 et 3. La chaîne de caractères $C = C[1] \dots C[N]$ est de longueur $N = 10000$.

La chaîne C est initialisée comme suit. On définit la suite récurrente d'entiers $(x_i)_{0 \leq i \leq N}$ par :
– x_0 est égal au numéro inscrit sur votre table d'examen,
– $x_i = ((2077 \times x_{i-1}) + 12345) \bmod 2^{16}$, pour i compris entre 1 et n.
Puis on remplace chaque x_i par la partie entière de $x_i/17$. Puis pour $1 \leq i \leq n$, on définit le i-ième caractère de la chaîne par $C[i] = x_i \bmod 4$.

(Remarque : notez bien que dans tout ce sujet les chaînes sont indicées à partir de 1, même si vous utilisez en machine des vecteurs ou des tableaux dont le premier élément est en position 0.)

(NB : Les numéros de table indiqués étaient 17, 19, 23, 29.)

Partie I. Préliminaires

Question I.1. Listez les 10 premières valeurs de la chaîne C : $C[1] \dots C[10]$.

Question I.2. Combien de caractères de C sont égaux à 0, 1, 2, 3 ?

Un *motif* M de longueur m est une chaîne de caractères $M[1] \dots M[m]$. C'est une *sous-chaîne* de C en position i si pour tout $k = 1, \dots, m$ on a $M[k] = C[i + k - 1]$.

La *fréquence* d'un motif M dans une chaîne C est le nombre de valeurs distinctes i telles que M soit une sous-chaîne de C en position i.

Question I.3. Quelle est la fréquence dans C du motif 0123 ?

Partie II. Notion de m-fréquence d'une chaîne

La *m-fréquence* d'une chaîne c de longueur n est égale à la fréquence du motif de longueur m le plus fréquent dans C. On cherche à déterminer la m-fréquence.

Question II.1. Décrivez brièvement un algorithme qui prend en entrée une chaîne c de longueur n et une longueur m et qui détermine la m-fréquence de c. Définissez parmi les opérations élémentaires effectuées par le programme celles qui sont les plus significatives (justifiez votre proposition !), et indiquez leur nombre en fonction de n et m.

(Aide : il pourra être utile d'encoder les différents motifs rencontrés, par exemple par des entiers, pour pouvoir mémoriser dans une table le nombre de leurs occurrences dans la chaîne.)

Question II.2. Quelle est la 1-fréquence de la chaîne C ?

Question II.3. Soit CC la chaîne formée des 10 premiers éléments de C. Quelle est sa 2-fréquence ? Sa 3-fréquence ? Sa 10-fréquence ? Dans chaque cas, précisez le(s) motif(s) le plus fréquent de la longueur considérée.

Question II.4. Quelle est 3-fréquence de C ? Sa 7-fréquence ? Dans chaque cas, précisez le(s) motif(s) le plus fréquent de la longueur considérée.

Question II.5. Pour quelle valeur maximale de m pouvez-vous calculer la m-fréquence de C en moins d'une minute ? Quelle est par exemple la 10-fréquence de C ? Sa 15-fréquence ? Sa 20-fréquence ?

À partir d'une certaine longueur tout motif est de fréquence 0 ou 1. La *longueur maximale de répétition* de la chaîne est la valeur maximale m pour laquelle il existe un motif de longueur m avec une fréquence d'au moins 2.

Question II.6. Décrivez brièvement un algorithme qui prend en entrée une chaîne c de longueur n et qui détermine la longueur maximale de répétition de c. Définissez parmi les opérations élémentaires effectuées par le programme celles qui sont les plus significatives (justifiez votre proposition !), et indiquez leur nombre en fonction de n.

Question II.7. Quelle est la longueur maximale de répétition de la chaîne CC de longueur 10 ? Quelle est celle de C ?

Partie III. Sous-chaînes sans ℓ-répétitions

Une sous-chaîne de longueur m d'une chaîne c est sans ℓ-répétition si sa longueur maximale de répétition est strictement inférieure à ℓ. Elle ne contient donc aucun motif de longueur ℓ répété.

Étant donné une chaîne c, on cherche à extraire des sous-chaînes sans ℓ-répétition de longueur maximale.

Question III.1. Décrivez brièvement un algorithme qui prend en entrée une chaîne c de longueur n et une longueur ℓ et qui détermine une sous-chaîne sans ℓ-répétition de longueur maximale. Définissez parmi les opérations élémentaires effectuées par le programme celles qui sont les plus significatives (justifiez votre proposition !), et indiquez leur nombre en fonction de n et de ℓ.

Question III.2. Donnez la longueur maximale d'une sous-chaîne de CC sans 2-répétition, et la décrire explicitement.

Question III.3. Donnez la longueur maximale d'une sous-chaîne de C sans 3-répétition.

Partie IV. Mariages de motifs

Deux motifs M et M' de même longueur m peuvent être *mariés* si pour tout $k = 1, \ldots .m$ on a $M[k] - M'[k] = 2 \bmod 4$.

Une chaîne contient un *mariage* de longueur m si elle contient deux sous-chaînes *disjointes*, chacune de longueur m et pouvant être mariées.

On pourra remarquer que ce problème est analogue à des questions pouvant survenir en génomique, car les mariages de motifs sont similaires à l'appariement de chaînes de nucléotides.

Question IV.1. Mêmes questions que précédemment, pour le calcul de la longueur maximale des mariages contenu dans C, en exprimant le nombre d'opérations en fonction de n.

Pour toute valeur ℓ on peut découper la chaîne en sous-chaînes disjointes de longueur au moins ℓ et réparties en un ensemble de paires mariées et un ensemble de motifs célibataires.

La *ℓ-auto-affinité* d'une chaîne est la valeur maximale possible du nombre de caractères appartenant aux sous-chaînes mariées.

Question IV.2. Mêmes questions que précédemment, pour le calcul de la ℓ-auto-affinité en exprimant le nombre d'opérations en fonction de n et ℓ. Donnez un encadrement de la 3-auto-affinité de C.