Devoir n°1 : Bio-Algorithmique
Recherche exacte de motifs

Indications

Travail à effectuer en binôme ou en monôme
A envoyer au plus tard le **dimanche 21 avril 2024** à l'adresse
recupspace@gmail.com
Le travail (compte rendu) doit être tapé dans un éditeur et enregistré au format pdf
Nom de fichier = « Devoir1-nomsEtudiants.pdf»
L'objet du mail : « Devoir1-BioALGO »

L'objectif de ce devoir est d'implémenter les algorithmes de recherche **exacte** de motifs dans un texte et d'analyser leur complexité théorique et expérimentale

1. Implémenter les algorithmes suivants pour rechercher toutes les occurrences de motifs dans un texte Tde longueur n, en prenant en compte le ***nombre de comparaisons*** effectuées :
- Algorithme de Boyer-Moore [BM] (recherche d'un seul motif M de longueur m)
- Algorithme de Aho-Corasick [AC] pour la recherche multiple (un ensemble de motifs S1, S2, …Sk)
- Algorithme de Rabin-Karp [RK] pour la recherche multiple d'un ensemble de motifs (S1, S2, …Sk), de même longueur, en utilisant les 2 variantes a. et b. suivantes
  - **a.** Utiliser 1 seule fonction de hachage de votre choix (voir TD).
  - **b.** Utiliser 3 fonctions de hachage (filtre de Bloom) de votre choix (voir TD).

2. Test et analyse des algorithmes [BM] et [RK]
- Faire un ensemble de tests de l'algorithme [BM] en reportant le ***nombre de comparaisons effectuées*** et le ***temps d'exécution*** *pour différents exemples avec différentes tailles du texte et du motif*. Montrer le meilleur des cas et le pire des cas.
- Faire un ensemble de tests de l'algorithme [RK] en reportant le ***nombre de comparaisons effectuées*** et le ***temps d'exécution*** *pour différents exemples,* ainsi que le nombre de faux positifs.
- Analyser la performance de l'algorithme **[RK]**, en termes du nombre de **faux positifs** induits en comparant les 2 variantes a/ et b/ (question 1). Conclure.

3. Comparaison des 2 algorithmes (**Aho-Corasick** [AC] et **Rabin-Karp** [RK])
- Faire un ensemble de tests en incluant le ***nombre de comparaisons effectuées*** et le ***temps d'exécution*** *des* deux algorithmes.

PS : pour les différents tests des algorithmes, il faudra reporter les résultats dans des tableaux en tenant compte des paramètres suivants : *la taille du texte, les tailles des motifs (ou la taille globale des motifs).*
- Tracer **les courbes** correspondant (diagrammes) aux résultats reportés dans les tableaux.
- Est-ce que les résultats de test des algorithmes précédents sont en accord avec la complexité théorique (voir cours) ?

4. L'algorithme de **Commentz-Walter** [CW] est un algorithme de recherche multiple (plusieurs motifs) qui se base sur les idées de l'algorithme de Aho-Corasick et de l'algorithme de Boyer-Moore.
   - Décrire le principe de l'algorithme de Commentz-Walter
   - Illustrer son déroulement sur un exemple de recherche de motifs S1, S2, S3 dans un texte T
   - Effectuez plusieurs tests pour comparer les performances de l'algorithme [CW] et [AC]
   - Analysez les résultats de test (procéder comme dans la question 2).

Remettre **un rapport** avec la structure suivante :

- Une page de garde (en tête, titre, noms des étudiants, ...)
- Une introduction dans laquelle on présente la bio-Informatique, l'intérêt de la recherche exacte de motifs et ses applications
- Le principe de chaque algorithme
- Les tableaux des tests,
- Les courbes (ou diagrammes)
- Une conclusion.
- **Une annexe** comportant le code source (bien commenté) des algorithmes