Recherche de séquences d'ADN.

Le but du projet est de concevoir un programme permettant la recherche rapide de séquences d'ADN. Ce procédé permet entre autre de diagnostiquer la présence d'un virus dans un organisme.

Une séquence d'ADN est représenté dans un fichier texte par une suite des caractères. Les caractères possibles sont uniquement : a, c, g, t et correspondent aux bases nucléique.

Le programme lit en entrée 2 fichiers :
- le premier contient la liste des séquences à rechercher. 1 ligne par séquence.
- le second contient la séquence à parcourir.

Le programme doit afficher sur la sortie standard (ou écrire sur un fichier) la liste des positions des séquences trouvés dans l'ordre ainsi que leur numéro (index de la ligne du fichier d'entrée).

Exemple :

Entrée à rechercher :
```
at
gat
tt
aca
```

Entrée à parcourir :
```
gatgattaca
```

Sortie :
```
0 1
1 0
3 1
4 0
5 2
7 3
```

Contraintes :
- Le nombre de séquence à rechercher varie de 1 à 1,000.
- Le nombre de caractère par séquence à rechercher varie de 10 à 10,000.
- La séquence à parcourir contient 100,000,000 caractères.
- La séquence à parcourir contient au moins une fois chaque séquence à rechercher (la plupart du temps, une et une seule fois)

Conseils :
- L'algorithme de Aho-Corasick permet de trouver toutes les sous-chaînes en une seule passe.
- Tester le code sur l'exemple avant les grandes données. Ne pas hésiter à écrire soi-même ses fichiers de tests.

Modalités :

Le projet est à faire en binôme ou monôme.
Le code doit être déposé sur **exam.ensiie.fr** sur le dépôt **iprf-2020-projet** avant le **22 avril 2020 à 23h59**, il est inutile d'écrire un rapport.
Une soutenance sera prévue pour la dernière séance où vous expliquerez la structure de votre code, vos choix et vos difficultés rencontrées.

La notation du projet se basera sur les critères de la validité des résultats, sur la vitesse d'exécution et sur la qualité de la soutenance.