Université de Nantes

UFR Sciences et Techniques

Algorithmique et programmation avancées pour les biologistes

Algorithmes et méthodes pour la Bio-Informatique

Extraction de motifs communs à plusieurs séquences biologiques

---

Séquence 1	ACTGGTTACCCTAGCATTCACG
Séquence 2	GTAGGTACTGAGGATAGACCGT

Ulysse Guyet
Jennifer Rondineau
Master 2 Bioinformatique

16 décembre 2015

Sommaire

1 Présentation du programme

Le programme s'utilise en ligne de commande. Il prend en entrée plusieurs paramètres :

1. Un nom de fichier FASTA contenant les séquences à étudier
2. La longueur du motif commun à extraire
3. Le nombre maximal d'erreurs autorisées entre occurrence et motif (distance de Levenshtein)
4. Le quorum, le pourcentage minimum de séquences qui doivent présenter une occurrence du motif
5. Éventuellement un nom de fichier dans lequel l'utilisateur pourra sauvegarder les résultats de l'extraction

```
./extraction_motif -f seq.fa -d 1 -q 0.5 -s sauvegarde.txt
```

L'argument -h permet d'afficher une aide à l'utilisation du programme:

```
moriarty@moriarty:~/Bureau/projetc$ ./extraction_motif -h

*****************************************************************
*Extraction de motifs communs à plusieurs séquences biologiques*
*****************************************************************

Usage : ./extraction_motif -f file.fa -d d_Levenshtein -q quorum

Options :
   -f,--filename_seq   FILE  Nom du fichier FASTA contenant les séquences à étudier
   -m,--longueur_motif   INT  Longueur du motif commun à extraire
   -d,--d_Levenshtein   INT  Le nombre maximal d'erreurs autorisées entre occurrence et motif (distance de Levenshtein)
   -q,--quorum   FLOAT   Le pourcentage minimum de séquences qui doivent présenter une occurrence du motif
   -s,--save   FILE   Nom du fichier dans lequel l'utilisateur veut sauvegarder les résultats de l'extraction
   -h,--help   FILE  Aide du programme
```

Après avoir entré les paramètres désirés à la suite du nom de l'exécutable, le menu suivant s'affiche:

```
moriarty@moriarty:~/Bureau/projetc$ ./extraction_motif -f seq_1.fa -m 6 -d 1 -q 0.5


        ********************************************************
        **********************   MENU   **********************
        ********************************************************

        1. afficher la liste des motifs
        2. choisir un motif et consulter a liste des occurrences de ce motif
        3. sortir du programme
```

1.1 Longueur du motif à extraire

L'option "-m" définit la longueur du motif commun que l'on désire extraire des séquences, cette longueur est de type entier. Si cette longueur dépasse la longueur maximale des séquences, le dictionnaire est alors vide.

1.2 La distance de Levenshtein

La distance de Levenshtein correspond au nombre maximal d'erreurs autorisées entre l'occurrence et le motif, cette variable est de type entier, et on l'a définit avec l'option "-d", par défaut cette distance est de 0. Les erreurs peuvent correspondre à une insertion, une délétion ou une substitution.

1.3 Le quorum

Le quorum correspond au pourcentage minimum de séquences qui doivent présenter une occurrence du motif. On définit cette variable par l'option "-q", elle est de type "float", et par défaut elle vaut 0.

1.4 La sauvegarde dans un fichier

Si l'utilisateur donne en paramètre l'option "-s" suivit d'un nom de fichier pour la sauvegarde, le programme enregistre dans ce fichier les résultats de l'extraction des motifs communs (seulement après consultation d'un motif).

Exemple de ligne de commande :

```
./extraction_motif -f seq.fa -q 0.5 -s sauvegarde.txt
```

Le fichier "sauvegarde.txt" se présente ainsi :

```
-------------------------- AAAAATG -------------------------------------
quorum : 0.666667
++++ seq 2 ++++
pos 519, ins 0, del 0, subst 0, last m
quorum : 0.666667
++++ seq 3 ++++
pos 378, ins 0, del 0, subst 0, last m
pos 863, ins 0, del 0, subst 0, last m
-------------------------- AAAACAA -------------------------------------
quorum : 0.666667
++++ seq 2 ++++
pos 175, ins 0, del 0, subst 0, last m
pos 259, ins 0, del 0, subst 0, last m
quorum : 0.666667
++++ seq 3 ++++
pos 108, ins 0, del 0, subst 0, last m
pos 844, ins 0, del 0, subst 0, last m
pos 1059, ins 0, del 0, subst 0, last m
```

Pour chaque motif, on sauvegarde le numéro de la séquence, la position, et les erreurs éventuelles (ins = insertion, del = délétion, subst = substitution, last = dernière opération réalisé (m = match, i = insertion, d = délétion, s = substitution).

2 Les structures de données utilisées

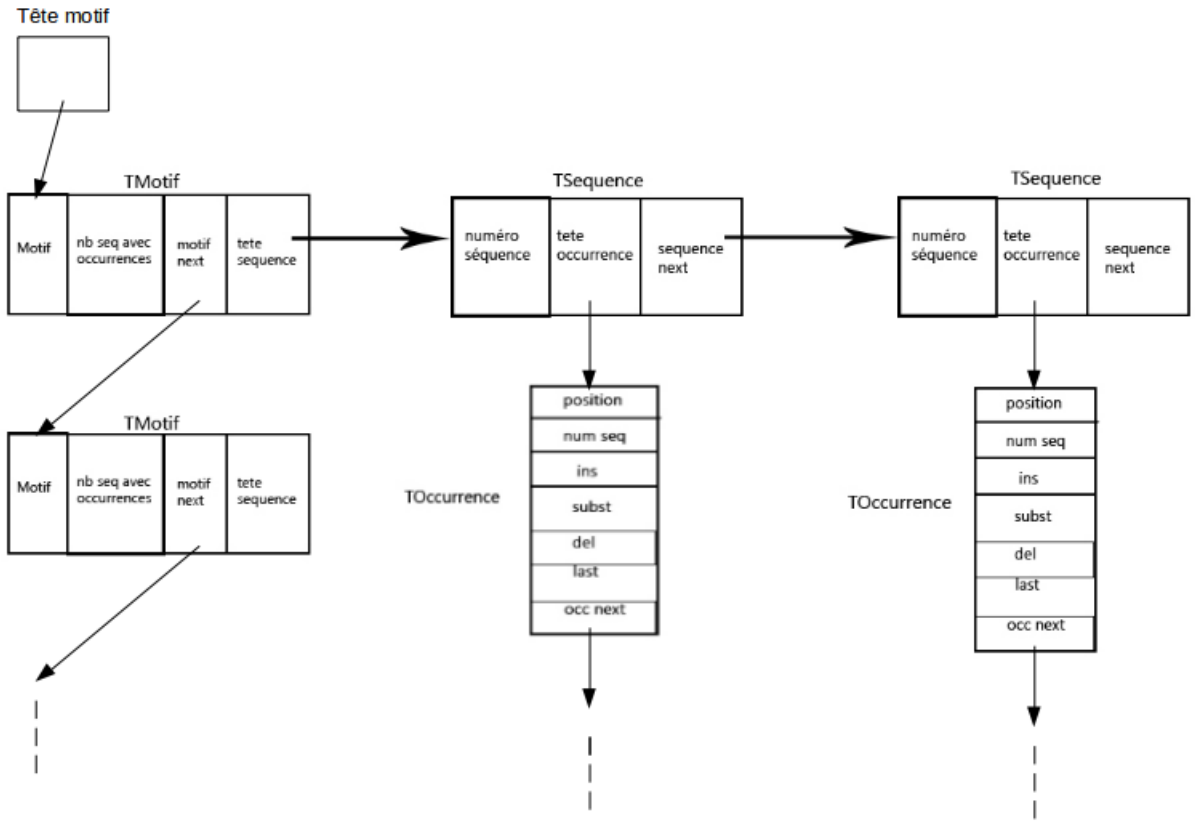


Figure 1: organisation des différentes structures

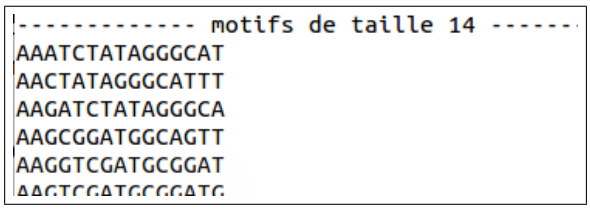
TeteMotif pointe vers une liste de motifs, qui pointe vers une liste de séquences et chaque séquence pointe vers une liste d'occurrences.

3 Tests réalisés

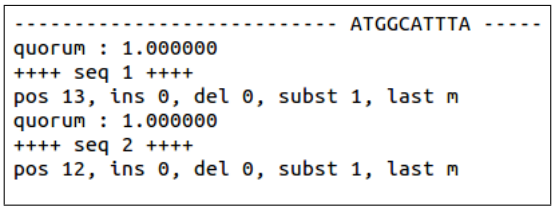
Nous avons d'abord effectué des tests sur des petites séquences, comme par exemple :

```
>Sequence1
AGGTCGATGCGGATGGCAGTTAA
>Sequence2
GGTAGATCTATAGGGCATTTA
```

Les figures ci-dessous montrent l'affichage de la liste des motifs (a) et le détail d'un motif en particulier (b). Nous avons également comparé nos résultats d'extractions avec un autre binôme. A partir d'un même fichier FASTA et des mêmes paramètres de recherche (quorum, distance de Levenshtein, longueur du motif), nous avons retrouvé les mêmes résultats.



(a) aperçu de l'affichage de la liste des motifs



(b) aperçu de l'affichage détail d'un motif

4 Les limites du programme

Si on désire réaliser une extraction de motifs communs à plusieurs séquences biologiques de taille importante, le temps de calcul devient assez conséquent. Plus les séquences sont longues, plus il faut du temps au programme pour extraire les motifs, surtout si on ajoute des erreurs possibles.

Sans erreur, en définissant simplement la taille du motif et le quorum, le programme s'exécute relativement vite (moins d'une seconde pour 3 séquences d'environ 600 nucléotides chacune).

En revanche pour ces même séquences, si on accepte seulement deux erreurs possibles, le programme mettra environ 14 min à extraire les motifs communs.

De même pour des séquences courtes :

```
>Sequence1
AGGTAGGAT
>Sequence2
AGGATTGA
```

Si on accepte 3 erreurs possibles sur un motif de longueur 7, et qu'on fixe le quorum à 50%, le programme met 5 minutes à s'exécuter et on obtient 10 033 motifs possibles.

5 Améliorations à apporter au programme

En ayant plus de temps, nous pourrions améliorer ce programme sur différents points :

1. L'enregistrement pourrait être amélioré, par exemple l'utilisateur pourrait enregistrer seulement les motifs qui l'intéresse.
2. L'affichage des motifs pourrait également être amélioré. On pourrait imaginer un menu qui propose à l'utilisateur de visualiser toutes les occurrences de motifs présentant un délétion, une substitution ou un match.
3. On pourrait rajouter des sécurités au niveau du getopt_long, par exemple, faire quitter le programme si l'utilisateur indique une chaîne de caractère au lieu d'un entier pour la distance de Levenshtein.