Identification de microARN

Les microARN sont des petites molécules d'ARN qui ont été découvertes dans les années 1990 chez les animaux et chez les plantes, et qui jouent un rôle essentiel dans le contrôle de l'expression des gènes. Chez l'homme, plus de 500 microARN ont été mis en évidence, et on sait maintenant que leur dysfonctionnement est associé à plusieurs maladies, comme le cancer. Le but de ce projet est d'écrire un programme qui permette d'identifier des microARN dans un génome.

Le fonctionnement des microARN

Les microARN sont issus d'ARN plus longs, les précurseurs de microARN, notés pré-miARN, qui font entre 70 et 100 nucléotides. Comme les ARN messagers, les pré-miARN sont codés dans les génomes et produits par transcription. Mais ils ne sont pas ensuite traduits en protéines. La séquence du pré-miARN se replie sur elle-même par appariement des nucléotides entre la première moitié et la deuxième moitié de la molécule, formant ainsi une structure en forme de tige-boucle. Les appariements se font suivant la complémentarité faible A-U, G-C et G-U.

Le pré-miARN est clivé (coupé) en microARN, d'une longueur de 22 nucléotides environ. Le clivage se fait dans la première partie de la tige boucle.

Par exemple, le pré-miARN ci-dessous fait 90 nucléotides et comprend 33 appariements. La première ligne est la séquence nucléique. La seconde ligne donne les informations sur les appariements. Chaque appariement est noté par une paires de parenthèses : “(” et “)”. Les nucléotides qui ne sont pas impliqués dans un appariement sont notés par des “.”.

```
UGCUUCCGGCCUGUUCCCUGAGACCUCAAGUGUGAGUGUACUAUUGAUGCUUCACACCUGGGCUCUCCGGGUACCAGGACGGUUUGAGCA
(((((((((.((((..(((.(((((.(((((.(((((((.(((.......))))))))))).)))).)))).)))...)))).)))...)))))
```

Il a une boucle terminale de 7 nucléotides (les points en vert au centre de l'image), et 12 boucles internes de 1 à 3 nucléotides (les autres points). Le microARN correspondant fait 21 nucléotides, et apparait entre les positions 15 et 35 (en rouge). Comment un microARN agit-il sur la régulation des gènes ? Le microARN et l'ARN messager sont tous les deux des molécules d'ARN simple brin. De ce fait, ces deux molécules peuvent s'apparier entre elles (comme peuvent s'apparier entre eux deux brins d'ADN). C'est que qu'on appelle l'hybridation. Un microARN s'apparie ainsi avec un ARN messager cible, et cette hybridation réprime la traduction de la protéine codée par l'ARN messager.

Dans le cas de l'exemple précédent, la séquence du microARN est UCCCUGAGACCUCAAGUGUGA, qui pourra s'hybrider sur le motif complémentaire AGGGACUCUGGAGUUCACACU. Tout ARN messager possédant ce précédent motif pourra s'hybrider avec le microARN exemple et voir sa traduction inhibée.

Le projet

Écrire un programme permettant de détecter tous les pré-microARN d'une séquence génomique.

La structure tige-boucle des pré-microARN à détecter est définie par les règles suivantes:

1) la longueur totale maximale est 100 nucléotides, dont au moins 48 sont impliqués dans des appariements (soit au moins 24 appariements);
2) tous les appariements sont emboités;
3) les appariements autorisés sont A-U, C-G et G-U;
4) les appariements apparaissent dans des groupes d'au moins trois nucléotides successifs appariés;
5) la boucle terminale est de longueur au plus 8 nucléotides, et les autres boucles sont de longueur au plus 3 nucléotides.

Le travail sera réalisé en sous tâches.

Sous tâche 1: Écrivez **vos tests d'abord !**
Écrivez un simulateur produisant aléatoirement des structures tige/boucle correspondant aux pré-microARN qui devront être détecter (les « vrais positifs » attendus), et entourant le microARN ainsi simulé de séquences aléatoires ne contenant pas la structure à détecter. Écrire dans un fichier la séquence simulée. Ce simulateur vous servira à produire les tests de votre programme de détection. Pensez donc à rendre le simulateur paramétrable, de manière à simplifier la mise au point, le débogage et la validation du programme de détection (sous tâche 2). Toujours pour les tests, il est intéressant que le simulateur donne la solution simulée et dont la détection est attendue.

Sous tâche 2 : **Recherche des pré-microARN**
Écrivez un algorithme efficace de recherche de toutes les tiges-boucles dans une séquence génomique, et implémenter le. Pensez à la programmation dynamique ! Lors du développement, testez votre programme en utilisant les séquences générée par le simulateur.

Sous tâche 3 : **Exploration de données biologiques**
Vous disposerez dès la prochaine séance d'un fragment de chromosome contenant des pré-microARN, combien et où sont-ils ? Vous disposerez de plus de la séquence d'un gène (ARN messager) pouvant s'hybrider avec un des microARN que vous aurez détecté. De quel microARN il s'agit, sachant que :

1) Sur les 7 première bases, la complémentarité entre le microARN et l'ARN messager est parfaite : A avec U, et G avec C,
2) le microARN fait entre 20 et 23 nucléotides,
3) il commence entre 10 et 15 nucléotides après la première position du pre-miARN.

Exemple :

```
ARN messager  ...  AGGGACUAUGG-GUUCAAGCCU ...
                   ||||||| ||| |||| || ||
microARN           UCCCUGAGACCUCAAG-UGUGA
```

A rendre

Vos sources commentées et un mini rapport d'un recto-verso au maximum, décrivant votre algorithme de recherche des pré-microARN (sous tâche 2), et vos résultats pour la sous tâche 3. Attention, pour la gestion de votre temps, passez au moins 50% sur la sous tâche 2 !