

Maîtrise d'informatique - Bioinformatique

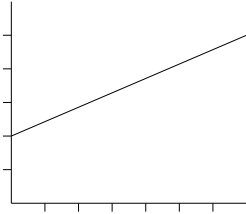
Examen de septembre 2001

Durée : 3 heures.

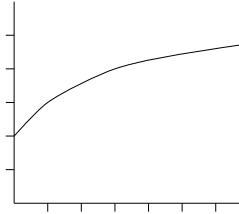
Documents autorisés. Calculatrices non autorisées

Exercice 1 : Alignement avec gaps concaves (4 points)

Quand on veut faire un alignement global entre deux séquences, les programmes usuels, tels que ceux que vous avez utilisés en TP, implémentent l'algorithme de Needleman et Wunsch avec des pénalités de gaps affines. Mais, le modèle le plus fin pour la prise en compte de gaps est d'affecter une pénalité du type $f(n)$ pour la création d'un gap de n nucléotides, où f est une fonction concave.¹



fonction affine



fonction concave

Écrire un algorithme qui permette de calculer la similarité globale de deux séquences d'ADN avec des pénalités de gaps concaves. On prendra le système de score suivant :

$$\begin{aligned} \alpha(x, y) &: \text{similarité entre les nucléotides } x \text{ et } y \\ f(n) &: \text{pénalité d'un gap de longueur } n \end{aligned}$$

Quelle est la complexité en temps, en espace ? À votre avis, pourquoi les programmes n'utilisent-ils pas des pénalités concaves, mais des pénalités affines ?

Exercice 2 : Le plus petit absent (6 points)

On considère le problème suivant :

Données : un ensemble de séquences $S = \{s_1, \dots, s_k\}$ définies sur un alphabet Σ .

Problème : trouver un mot u de longueur minimale qui ne soit facteur d'aucune des séquences de S .
 u est appelé un plus petit absent de S .

Question 1. Pour l'exemple $S = \{aab, baaa, baab\}$ sur l'alphabet $\{a, b\}$, quel est le plus petit absent (ou l'un des plus petits absents) ?

Question 2. Construire l'arbre des suffixes généralisé de S . Comment déterminer les plus petits absents à partir de cet arbre ?

Question 3. Donner un algorithme qui permette de trouver le plus petit absent pour un ensemble S et un alphabet Σ quelconques. Vous devez trouver un algorithme en temps $O(|\Sigma| \times |S|)$.

On considère maintenant le problème voisin suivant, qui est à la base de la méthode de séquençage par *primer walking* :

Données : un ensemble de séquences $S = \{s_1, \dots, s_k\}$ définies sur un alphabet Σ , une séquence α sur l'alphabet Σ et une position i de α .

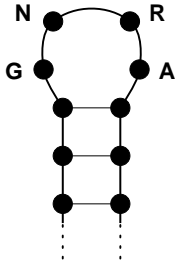
Problème : trouver le plus petit mot absent de S et présent dans α à la position i .

¹Une fonction *concave* est une fonction dont la dérivée première est positive, et la dérivée seconde est négative, comme le logarithme par exemple. Vous n'avez pas besoin de cette définition pour faire l'exercice.

Question 4. Décrire un algorithme permettant de résoudre ce problème.

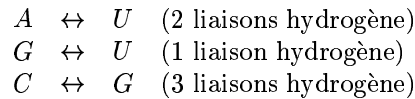
Exercice 3 : Les boucles GNRA (4 points)

Dans les séquences d'ARN, on trouve des petits motifs structuraux, appelés des boucles GNRA, définis comme suit



- le motif commence par une tige, c'est-à-dire une succession de nucléotides appariés,
- la tige encadre une boucle de longueur exactement 4, constituée dans l'ordre d'une guanine **G**, d'un nucléotide quelconque (**N=A, G, C** ou **U**, avec la même probabilité), d'une adénine ou d'un uracile (**R =A** ou **U**) et d'une adénine **A**.

Nous vous rappelons que dans l'ARN, les appariements possibles sont



On fait l'hypothèse que pour être suffisamment stable une tige d'une boucle GNRA doit comporter au moins 5 liaisons hydrogène.

Question 1. Serait-il adéquat de modéliser les boucles GNRA à l'aide d'un Modèle de Markov ? Pourquoi ?

Question 2. Proposez un *modèle* alternatif permettant de détecter les boucles GNRA.

Exercice 4 : Pour finir (6 points)

Toutes les questions sont indépendantes.

Question 1. Donner un modèle de Markov pour l'alignement multiple ci-dessous.

```

G T T A C G - T A A
T T A A C G C T A A
G T - - C G - T A C
G T A A C G - T A A
    
```

Question 2. Dans Genbank, BLAST trouve le motif ACCTGCGTAGATTCA avec une E-value égale à 0.75. Pour la même banque de données, quelle serait la E-value pour l'occurrence du motif ACCTGCGTAGATTCAG ?

Question 3. Pour comparer deux séquences, au lieu de chercher à maximiser la *similarité*, on peut minimiser la *distance*, en comptant le nombre de différences. Le système de score est alors du type :

$$\begin{aligned}
 \text{Indel}(x) &> 0, \text{ coût de l'insertion ou de la délétion du nucléotide } x, \\
 \text{Distance}(x,y) &> 0 \text{ si } x \neq y, \text{ et } 0 \text{ si } x = y.
 \end{aligned}$$

Donner l'algorithme pour l'alignement global de deux séquences en minimisant la distances entre celles-ci. Que pensez-vous du problème de l'alignement local ?