# Option Informatique en Spé MP et MP*

## Devoir à rendre après les vacances de la Toussaint

Structure secondaire de l'ARN de transfert

**Résumé**

On regroupe sous le nom générique de *molécules de la vie* l'ADN, les diverses sortes d'ARN et les protéines. Leur étude fait intervenir de multiples spécialistes : chimistes, biochimistes, généticiens... Elle pose d'intéressants problèmes aux informaticiens et aux combinatoristes.

On observe ici quelques propriétés de la *structure secondaire* des ARN de transfert ; ces macromolécules interviennent dans la synthèse des protéines.

Dans un premier temps, on s'intéresse au nombre $S(n)$ de ≪modèles≫ possibles de structures secondaires de longueur $n$ ; on donne un majorant de $S(n)$.

On étudie ensuite diverses représentations d'un tel modèle : au moyen d'un arbre, au moyen d'une chaîne de caractères.

La dernière partie propose le calcul d'une structure secondaire optimale (en un sens défini dans le texte), au moyen d'une technique classique de programmation dynamique.

*Veuillez rédiger chaque partie sur une copie séparée.*

## Table des matières

**1 Dénombrement des modèles de structures secondaires** **3**

**2 Représentation d'un modèle par un arbre** **4**

**3 Représentation linéaire d'un modèle** **5**

**4 Optimisation d'une structure secondaire** **5**

## Notations, définitions, et mises en garde

▶ Un ARN de transfert est une macro-molécule formée (pour l'essentiel) d'une succession de bases purines (adénine et guanine) et pyrimidines (cytosine et uracile), deux bases consécutives étant reliées par une *liaison phosphodiester*. On peut considérer cette molécule comme un mot sur l'alphabet $\mathcal{B} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$ ; ce mot représente la *structure primaire* de l'ARN.

▶ Cette molécule, qui devrait avoir l'apparence d'un fil, a tendance à se replier sur elle-même, car certaines paires de bases non adjacentes sont reliées par une *liaison hydrogène*. L'ensemble de ces liaisons constitue la *structure secondaire* de l'ARN. La description que nous venons de faire est volontairement simplifiée, mais elle suffit pour nos besoins. La figure 1 donne un exemple de telle structure ; les liaisons primaires apparaissent en traits fins, tandis que les liaisons secondaires apparaissent en traits plus épais.

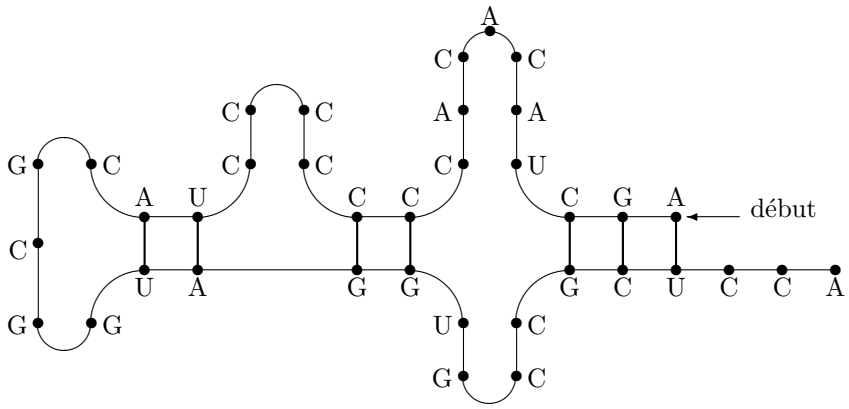


Figure 1: un exemple (hypothétique) de structure secondaire

*On peut admettre un résultat, à condition de le signaler clairement ; en tout état de cause, il est demandé de rédiger les questions dans l'ordre de l'énoncé. Les programmes devront être concis, et suffisamment documentés pour être compréhensibles. L'emploi de références est interdit. Les questions marquées ⋆⋆⋆ sont, à mon sens, plus délicates.*

# 1 Dénombrement des modèles de structures secondaires

▶ On ne se préoccupe pas, pour l'instant, de la nature précise des bases qui forment la structure primaire de l'ARN de transfert.

▶ Un *modèle de structure secondaire* de longueur $n$ est une involution $s$ de l'intervalle discret $[\![1, n]\!]$ vérifiant les deux propriétés suivantes :

**(1)** si $i$ n'est pas un point fixe de $s$, alors $|s(i) - i| \geqslant 2$.

**(2)** soient $i$ et $j$ deux éléments de $[\![1, n]\!]$ tels que $i < s(i)$, $j < s(j)$ et $i < j$ ; on a alors soit $s(i) < j$ soit $s(j) < s(i)$.

Définissons une *liaison secondaire* du modèle $s$ : c'est un couple $(i, s(i))$ tel que $i < s(i)$. La première propriété traduit le fait qu'une liaison secondaire relie deux bases non adjacentes ; la deuxième propriété traduit l'hypothèse simplificatrice suivante : la structure secondaire de l'ARN de transfert est représentable par un graphe planaire. Ceci revient à dire que les liaisons secondaires peuvent être imbriquées, mais non enchevêtrées.

▶ Dans la suite, nous utiliserons simplement le mot *modèle* à la place de l'expression *modèle de structure secondaire*. Un modèle $s$ de longueur $n$ pourra être représenté par la liste $(s(1), s(2), \ldots, s(n))$ des images par $s$ des éléments de $[\![1, n]\!]$. Le modèle *banal* est celui dans lequel il n'y a aucune liaison. La figure 2 donne deux représentations différentes d'un même modèle.

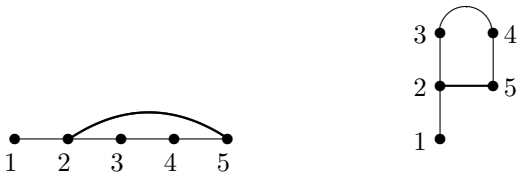


Figure 2: deux représentations du modèle $(1, 5, 3, 4, 2)$

***Question 1*** • Soient $s$ un modèle de longueur $n$, et $(i, s(i))$ une liaison secondaire de $s$. Montrez que l'intervalle $[\![i, s(i)]\!]$ est stable par $s$.

▶ On peut alors considérer le modèle de longueur $p = s(i) - i - 1$ *induit* par $s$ sur l'intervalle $[\![i+1, s(i)-1]\!]$ : c'est l'application $t$ de l'intervalle $[\![1, p]\!]$ dans lui-même définie par $t(k) = s(k + i) - i$.

▶ Pour les fonctions à écrire en Caml, un modèle $s$ sera représentée par un vecteur d'entiers contenant, dans cet ordre, les valeurs $s(1), s(2), \ldots, s(n)$. Ne pas oublier qu'en Caml, l'indexation des éléments d'un vecteur commence à 0.

***Question 2*** • Énumérez les modèles de longueur 4, puis ceux de longueur 5.

***Question 3*** • Soient $n \geqslant 1$ et $s$ une application de l'intervalle $[\![1, n]\!]$ dans lui-même, autre que l'identité. Soit $i$ le plus petit indice tel que $s(i) \neq i$. Montrez que $s$ est un modèle ssi les deux conditions suivantes sont satisfaites :

1. $s(i) > i + 1$ ;
2. $s$ induit un modèle sur chacun des intervalles $[\![i + 1, s(i) - 1]\!]$ et $[\![s(i) + 1, n]\!]$.

***Question 4*** • Rédigez en Caml une fonction de type `int vect -> bool` qui indique si un vecteur d'entiers représente un modèle.

▶ On note $S(n)$ le nombre de modèles de longueur $n$.

***Question 5*** • Dressez la liste des valeurs de $S(n)$ pour $n \leqslant 5$.

***Question 6*** • Établissez la relation

$$S(n+2) = S(n+1) + S(n) + \sum_{1 \leqslant i < n} S(i)S(n-i)$$

***Question 7*** • Rédigez en Caml une fonction de type `int -> int` calculant la valeur de $S(n)$ pour $n$ donné. Le coût $c(n)$ du calcul de $S(n)$ ne devra pas être exponentiel en $n$ (une analyse de ce coût serait d'ailleurs bienvenue).

***Question 8*** • Rédigez en Maple une fonction calculant $S(n)$.

***Question 9*** • Proposez un minorant simple de $S(n)$, montrant que la croissance de cette suite est exponentielle.

***Question 10*** ⋆⋆⋆ • Justifiez (ou admettez) la majoration $S(n) \leqslant \dfrac{3^n}{n(n+1)}$.

***Question 11*** Que pouvez-vous en déduire, concernant le rayon de convergence de la série entière $\sum_{n \geqslant 1} S(n)z^n$ ?

▶ On note $\varphi(z) = \sum_{n \geqslant 1} S(n)z^n$ lorsque cette série converge.

***Question 12*** ⋆⋆⋆ Montrez que $\varphi(z)$ vérifie une équation du second degré.

▶ Soient $i$ et $j$ deux indices tels que $j - i \geqslant 2$, ni $i$ ni $j$ ne sont points fixes de $s$, et tout $k \in [\![i+1, j-1]\!]$ est point fixe de $s$. On dit que $(i, j)$ est une *boucle finale* si $i$ et $j$ sont images l'un de l'autre par $s$. On dit que $(i, j)$ est un *bulbe* si $s(i) = s(j) + 1$. On dit que $(i, j)$ est une *boucle interne* si $s(i) - s(j) > 1$ et si tout $k \in [\![s(j)+1, s(i)-1]\!]$ est invariant par $s$. Dans les autres cas de figure, on dit que $i$ et $j$ participent à une *jonction*.

***Question 13*** • Énumérez les boucles (internes et finales), les bulbes et les jonctions du modèle représenté à la figure 1 ; la première base est visée par la flèche.

***Question 14*** • Montrez que tout modèle comportant au moins une liaison comporte au moins une boucle finale.

▶ Un modèle est une *épingle à cheveux* s'il comporte exactement une boucle finale, éventuellement des boucles internes et/ou des bulbes, mais aucune jonction.

***Question 15*** • Combien existe-t-il d'épingles à cheveux de longueur $n$ ?

## 2 Représentation d'un modèle par un arbre

▶ *Les 3/2 ne sont pas obligés de traiter cette partie.*

▶ Soit $s$ un modèle de longueur $n$. Un point fixe $k$ de $s$ est *visible* s'il n'existe aucune liaison secondaire $(i, s(i))$ telle que $i < k < s(i)$. De la même façon, une liaison secondaire $(k, s(k))$ est *visible* s'il n'existe aucune liaison secondaire $(i, s(i))$ telle que $i < k$ et $s(k) < s(i)$.

▶ On associe à $s$ un arbre de la façon suivante : les fils de la racine sont les éléments visibles, dans l'ordre de leurs indices. Chaque point fixe visible est une feuille ; chaque liaison secondaire visible $(k, s(k))$ est un nœud, racine de l'arbre représentant le modèle induit par $s$ sur l'intervalle $[\![k+1, s(k)-1]\!]$. La figure 3 illustre cette idée

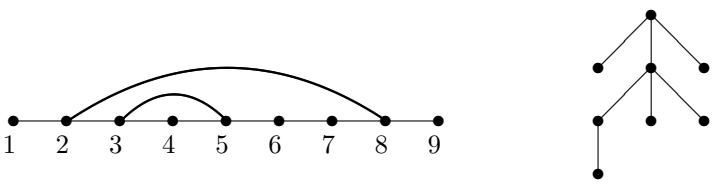


Figure 3: un modèle et l'arbre qui lui est associé

***Question 16*** • Montrez que l'on définit *effectivement* une fonction, notée $\tau$ dans la suite.

***Question 17*** • Construisez l'arbre associé au modèle dessiné à la figure 1.

***Question 18*** • Exprimez le nombre de nœuds et le nombre de feuilles de l'arbre associé à un modèle $s$, en fonction de la longueur $|s|$ et du nombre $\ell(s)$ de liaisons secondaires de ce modèle.

***Question 19*** • Montrez que l'application $\tau$ est injective.

▶ On définit le type Caml `type arbre = F | N of arbre list;;`.

***Question 20*** • Rédigez en Caml une fonction :

```
tau : int vect -> arbre
```

qui calcule l'arbre associé à un modèle, ce dernier étant représenté par un vecteur.

***Question 21*** • Rédigez en Caml une fonction :

```
inv_tau : arbre -> int vect
```

inverse de la précédente.

# 3 Représentation linéaire d'un modèle

▶ À un modèle $s$ de longueur $n$, on associe sa *représentation linéaire* : c'est un mot $\psi(s)$ sur l'alphabet $\mathcal{R} = \{x, g, d\}$ défini comme suit : notons $j = s(1)$ ; si $j = 1$, alors $\psi(s) = x\psi(t)$ où $t$ est le modèle induit par $s$ sur l'intervalle discret $[\![2, n]\!]$ ; si $j = n$, alors $\psi(s) = g\psi(t)d$ où $t$ est le modèle induit par $s$ sur l'intervalle discret $[\![2, n-1]\!]$ ; enfin, si $1 < j < n$, alors $\psi(s) = g\psi(t)d\psi(u)$ où $t$ est le modèle induit par $s$ sur l'intervalle discret $[\![2, j-1]\!]$ et $u$ le modèle induit par $s$ sur l'intervalle discret $[\![j+1, n]\!]$. Par exemple, au modèle de la figure 3 on associe le mot $xggxdxxdx$.

***Question 22*** • Montrez que ceci définit *effectivement* une fonction.

***Question 23*** • On note $|w|_a$ le nombre d'occurrences d'une lettre $a$ dans un mot $w$. Exprimez $\big|\psi(s)\big|_x$, $\big|\psi(s)\big|_g$ et $\big|\psi(s)\big|$ en fonction de la longueur $|s|$ et du nombre $\ell(s)$ de liaisons secondaires de $s$.

***Question 24*** • Calculez la représentation linéaire du modèle dessiné à la figure 1.

▶ *Les 3/2 ne sont pas obligés de traiter la fin de cette partie.*

***Question 25*** • Rédigez en Caml une fonction :

```
psi : arbre -> int string
```

qui, à un arbre $t$, associe la représentation linéaire du modèle décrit par $t$.

***Question 26*** • Montrez que l'application $\psi$ est injective.

***Question 27*** • Rédigez en Caml une fonction :

```
inv_psi : string -> arbre
```

qui, à une chaîne $w$ associe l'arbre décrivant le modèle dont $w$ est la représentation linéaire, ou déclenche une exception si $w$ n'est pas la représentation linéaire d'un modèle.

# 4 Optimisation d'une structure secondaire

▶ Jusqu'ici, la nature précise des bases constituant l'ARN de transfert n'est pas intervenue. Soit $m$ un mot de longueur $n$ sur l'alphabet $\mathcal{B}$ ; une *structure secondaire* sur $m$ est un modèle de longueur $n$ vérifiant la propriété suivante :

**(3)** pour toute liaison secondaire $(i, s(i))$ apparaissant dans $s$, le couple $\big(m_i, m_{s(i)}\big)$ appartient à l'ensemble $\big\{(\mathrm{A}, \mathrm{U}); (\mathrm{U}, \mathrm{A}); (\mathrm{C}, \mathrm{G}); (\mathrm{G}, \mathrm{C})\big\}$.

Cette propriété traduit le fait que seules certaines paires de bases peuvent établir une liaison hydrogène. On vérifiera que la figure 1 représente bien une structure secondaire.

***Question 28*** • Énumérez toutes les structures secondaires possibles sur le mot $m =$ ACAGGUC.

▶ On se propose de décrire un algorithme qui, étant donné un mot $m$ de longueur $n$ sur l'alphabet $\mathcal{B}$, détermine le nombre maximal $\lambda(m)$ de liaisons d'une structure secondaire sur $m$. Pour $1 \leqslant i \leqslant j \leqslant n$, on note $\ell(i,j) = \lambda(m[i..j])$ où $m[i..j]$ désigne le mot $m_i m_{i+1} \dots m_{j-1} m_j$.

***Question 29*** • Pour $1 \leqslant i \leqslant j < n$, montrez que $\ell(i, j+1)$ est la plus grande des deux valeurs $\ell(i,j)$ et $\sigma(i,j)$ où

$$\sigma(i,j) = \max_{i \leqslant k < j} \rho(k, j+1)\big(\ell(i,k-1) + 1 + \ell(k+1,j)\big)$$

et $\rho(p,q)$ est égal à 1 si les bases $m_p$ et $m_q$ peuvent établir une liaison secondaire, et à 0 sinon.

***Question 30*** • En déduire un algorithme de calcul de $\lambda(m)$.

***Question 31*** • Exprimez la complexité de cet algorithme en fonction de $n$. Plus précisément, vous déterminerez un exposant $\alpha$ tel que le coût (en nombre de consultations d'éléments de vecteurs) de l'application de cet algorithme à une structure de longueur $n$ soit un $\mathcal{O}(n^\alpha)$.

***Question 32*** • Rédigez en Caml une fonction :

```
lambda : string -> int
```

qui réalise le calcul de la fonction $\lambda$.

***Question 33*** • Rédigez en Caml une fonction :

```
optimise : string -> int vect
```

qui détermine une structure secondaire optimale, c'est-à-dire dont le nombre de liaisons secondaires est maximal.

## FIN