Agrégation externe d'informatique *Épreuve de modélisation*
Session N/A (sujet 0)

Énumération et algorithmique des structures secondaires

Le jury attend une présentation de 35 minutes s'appuyant sur ce sujet, en forme de cours, pédagogique, structurée et évitant la paraphrase. Le texte se conclut par des pistes de réflexion facultatives dont vous pouvez vous saisir ; au-delà de ces pistes proposées, toute initiative personnelle pertinente est appréciée. Vous n'êtes pas obligé de traiter le texte dans son intégralité, mais si seule une partie est traitée elle doit l'être de manière particulièrement approfondie.

L'exposé doit intégrer une (ou plusieurs) illustrations informatiques. Il doit également contenir une discussion autour d'une dimension éthique, sociétale, environnementale, économique ou juridique en lien avec le texte ; une des pistes de réflexion est spécifiquement conçue à cet effet.

1 Contexte

Depuis une vingtaine d'années, les moyens d'observation sont devenus suffisamment performants pour étudier le contenu d'une cellule au niveau moléculaire. On cherche dorénavant à exploiter les informations collectées pour comprendre finement des fonctions biologiques complexes telles que la reproduction, l'auto-réparation ou la capacité d'adaptation au milieu de vie.

Au cœur de la cellule, on trouve de longues molécules d'ADN qui forment les génomes. C'est là que résident toutes les données nécessaires à la maintenance et à la reproduction de ces cellules. Néanmoins, on sait à présent que la molécule d'ADN n'est pas le seul lieu d'intérêt pour la compréhension du vivant.

La molécule d'ARN ressemble à l'ADN, mais est plus dynamique. Son rôle dans la cellule est davantage de l'ordre de la traduction que du stockage. L'ARN est un simple brin constitué de *bases nucléiques* (dans la suite nous écrirons simplement *bases*), qui sont l'adénine (notée A), la cytosine (notée C), la guanine (notée G), et l'uracile (notée U) ; on modélise donc usuellement une séquence d'ARN comme un mot sur l'alphabet $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$ – ce que nous ferons également dans la suite. Ce brin peut se replier sur lui-même dans l'espace en associant les lettres A−U et G−C, en des repliements qui donnent des formes très variées. Malgré la puissance des outils d'observation actuels, l'étude des structures de l'ARN en 3 dimensions dans une cellule reste très compliquée. Les méthodes expérimentales actuelles s'appuient donc sur des modèles formels.

Nous proposons ici d'étudier des propriétés d'une structure intermédiaire que l'on appelle structure secondaire, composée de la séquence elle-même et d'un ensemble d'appariements entre ses bases, soumis à des contraintes.

Nous précisons dans un premier temps la définition des structures secondaires avant d'étudier un algorithme de prédiction de la structure secondaire la plus probable pour une séquence d'ARN donnée, ainsi qu'une approche permettant de générer des structures secondaires. Enfin, dans une dernière partie, nous nous intéressons au nombre moyen de structures secondaires possibles pour une séquence ARN fixée.

1.1 Séquences d'ARN et appariements

Une séquence d'ARN est une suite finie de lettres A, C, G et U. Si w est une séquence d'ARN de longueur n, on note w_i sa $i$ème lettre ($1 \leq i \leq n$). La lettre A est la *lettre complémentaire* de U, U est la lettre complémentaire de A et les lettres C et G sont également complémentaires l'une de l'autre. Dans une séquence d'ARN w de longueur n, on dit qu'un couple d'entiers (i, j) tel que $1 \leq i < j \leq n$ est un *appariement* si w_i est la lettre complémentaire de w_j. Deux appariements (i_1, j_1) et (i_2, j_2) dans une séquence d'ARN w sont *disjoints* si les quatre indices i_1, i_2, j_1 et j_2 sont tous différents les uns des autres. On dit que deux appariements *se croisent* ou qu'ils sont croisés si $i_1 \in [i_2, j_2]$ et $j_2 \in [i_1, j_1]$, ou si $i_2 \in [i_1, j_1]$ et $j_1 \in [i_2, j_2]$.

Par exemple, pour la séquence AAGAACCGUUGAAAC et l'ensemble d'appariements $\mathcal{A} = \{a_1, a_2, a_3, a_4\}$ avec $a_1 = (2, 10)$, $a_2 = (3, 6)$, $a_3 = (5, 9)$, $a_4 = (11, 15)$, les appariements a_2 et a_3 se croisent. Ce n'est le cas d'aucune autre paire d'appariements dans l'ensemble $\mathcal{A}$. Tous les appariements de $\mathcal{A}$ sont deux à deux disjoints. Ce ne serait pas le cas, par exemple, si on y ajoutait l'appariement $a_5 = (3, 7)$, qui croise l'appariement $a_2 = (3, 6)$.

1.2 Structures secondaires

Lorsque l'on considère une séquence d'ARN associée à un ensemble d'appariements disjoints et sans aucun croisement sur cette séquence, on parle d'une *structure secondaire*. Plusieurs structures secondaires sont possibles pour une même séquence d'ARN. Ainsi, on peut construire des structures secondaires différentes sur la séquence AAGAACCGUUGAAAC, en lui associant les ensembles d'appariements suivants : $\mathcal{A}_1 = \{(2, 10), (3, 6), (11, 15)\}$, $\mathcal{A}_2 = \{(3, 15), (4, 10), (5, 9)\}$ ou encore $\mathcal{A}_3 = \{(3, 6), (7, 8), (9, 13), (10, 12)\}$.

2 Algorithmique des structures secondaires

Posons $X = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$. Étant donné un mot de X^*, le nombre de structures secondaires possibles pour cette séquence peut être très élevé – nous y reviendrons dans la dernière partie et verrons qu'il est en moyenne exponentiel en la longueur de la séquence. Toutefois, dans la nature, une séquence donnée se replie en un nombre limité de structures, selon les lois de la thermodynamique.

Les algorithmes classiques de prédiction de la ou des structures secondaires d'une séquence sont fondées sur la recherche d'une ou plusieurs structures d'énergie libre minimale – l'énergie d'une structure est modélisée par une fonction dont certains paramètres ont été obtenus par des expérimentations biologiques. Le problème de prédiction de la structure secondaire d'une séquence d'ARN donnée consiste, dans sa version la plus simple, à trouver une structure d'énergie minimale parmi toutes

les structures secondaires théoriquement possibles. Il s'agit donc d'un problème d'optimisation combinatoire.

Nous allons nous intéresser à l'un des premiers algorithmes de prédiction à avoir été publié, dans les années 1970. Il est fondé sur un modèle d'énergie trop simple pour aboutir à de bonnes prédictions. Cependant, il a donné naissance à toute une famille d'algorithmes de même nature, fondés sur un modèle plus réaliste, et qui sont aujourd'hui parmi les meilleurs connus pour prédire les structures secondaires d'ARN.

2.1 Maximisation du nombre de paires

Le modèle d'énergie est le suivant : l'énergie d'une structure est égale à l'opposé de son nombre d'appariements. Trouver une structure d'énergie minimale revient donc à trouver une structure qui a un nombre maximal d'appariements.

On suppose[1] dans la suite que deux bases complémentaires peuvent s'apparier quelle que soit leur distance dans la séquence.

Une stratégie naïve consiste à énumérer toutes les structures secondaires et à chercher la structure d'énergie minimale ; toutefois, comme indiqué plus haut, ce nombre de structures secondaires est en général exponentiel et cette stratégie est condamnée pour n grand.

Notons $\theta_{i,j}$ le nombre maximal d'appariements d'une structure secondaire sur la séquence d'ARN $w_i \dots w_j$. On peut alors calculer le nombre maximal d'appariements pour w en utilisant les formules suivantes :

$$\theta_{i,i-1} = 0 \quad \text{pour } 2 \leq i \leq n \ , \tag{1}$$

$$\theta_{i,i} = 0 \quad \text{pour } 1 \leq i \leq n \ . \tag{2}$$

pour tout $1 \leq i < j \leq n$,

$$\theta_{i,j} = \max\left(\theta_{i+1,j}, \theta_{i,j-1}, \theta_{i+1,j-1} + \delta_{i,j}, \max_{i<k<j}[\theta_{i,k} + \theta_{k+1,j}]\right) \tag{3}$$

avec

$$\delta_{i,j} = \begin{cases} 1 & \text{si } w_i \text{ est la lettre complémentaire de } w_j \ , \\ 0 & \text{sinon} \quad . \end{cases} \tag{4}$$

On remplit alors pas à pas une matrice avec tous les $\theta_{i,j}$. La matrice associée à la séquence CCCUUUAGG est donné dans la Table 1.

Les algorithmes les plus récents (et les plus efficaces) de prédiction de structure secondaire sont inspirés de cette approche originelle. Cependant les modèles d'énergie sont bien plus réalistes et donnent lieu à des relations de récurrence plus complexes. D'autre part, ils ne recherchent pas nécessairement la ou une structure d'énergie minimale, mais ils ont une approche probabiliste en recherchant un ensemble de structures qui sont les plus probables dans le modèle d'énergie donné.

[1] Ceci est également une simplification car dans la réalité, deux bases ne peuvent s'apparier que si au moins trois autres bases les séparent.

-	C	C	C	U	U	U	A	G	G
C	0	0	0	0	0	0	1	2	3
C	0	0	0	0	0	0	1	2	3
C	-	0	0	0	0	0	1	2	2
U	-	-	0	0	0	0	1	1	1
U	-	-	-	0	0	0	1	1	1
U	-	-	-	-	0	0	1	1	1
A	-	-	-	-	-	0	0	0	0
G	-	-	-	-	-	-	0	0	0
G	-	-	-	-	-	-	-	0	0

Table 1: table finale pour la séquence CCCUUUAGG

2.2 Optimisations

Les séquences ARN peuvent être très longues et leur manipulation demande un espace de stockage de l'ordre du Gigaoctet. Elles appellent naturellement à une représentaton efficace. Par ailleurs, la complexité de calcul des structures secondaires impose l'utilisation du parallélisme.

2.2.1 Encodage efficace

Les séquences ARN étant composées de lettres, il est possible d'utiliser le codage standard ASCII. Toutefois, comme seuls quatre symboles sont possibles, on peut coder chaque base w_i sur 2 bits a_i et b_i. Une solution directe est d'encoder les 4 bases en utilisant les quatres valeurs possibles à deux bits et représenter la séquence de manière contiguë base par base. Cette approche divise par 4 l'espace de stockage nécessaire. Ce codage permet l'expression de $\delta_{i,j}$ comme une fonction booléenne de a_i, b_i, a_j, b_j.

2.2.2 Comparaison de séquences ARN

On s'intéresse ici au problème d'étudier l'appariement global de deux séquences $w_i, w'_i, 1 \leq i \leq k$ de même longueur k en comparant, pour chaque position, w_i et w'_i pour voir si les deux bases correspondantes sont complémentaires l'une de l'autre (i.e., si le couple (w_i, w'_i) est l'un de $(\mathrm{A}, \mathrm{U}), (\mathrm{U}, \mathrm{A}), (\mathrm{C}, \mathrm{G}), (\mathrm{G}, \mathrm{C})$).

Il est possible de traiter ce problème en utilisant des instructions logiques bit-à-bit pour comparer plusieurs positions en parallèle, en utilisant un codage adapté.

2.3 Grammaires

On peut engendrer une structure secondaire en utilisant une *grammaire* non contextuelle. La grammaire $\mathcal{G}_1$ ci-dessous a 14 règles de production et un seul non-terminal, S, qui est aussi le symbole de départ. Ici et dans la suite de cette partie, pour éviter toute confusion entre $\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}$ (qui sont des symboles terminaux) et les symboles non-terminaux, on écrit ces derniers en gras.

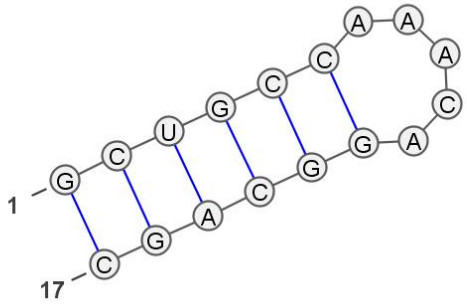
$$
\begin{array}{lcll}
\boldsymbol{S} & \to & \mathrm{A}\boldsymbol{S} \,|\, \mathrm{C}\boldsymbol{S} \,|\, \mathrm{G}\boldsymbol{S} \,|\, \mathrm{U}\boldsymbol{S} & \text{premier caractère non apparié} \\
\boldsymbol{S} & \to & \boldsymbol{S}\mathrm{A} \,|\, \boldsymbol{S}\mathrm{C} \,|\, \boldsymbol{S}\mathrm{G} \,|\, \boldsymbol{S}\mathrm{U} & \text{dernier caractère non apparié} \\
\boldsymbol{S} & \to & \mathrm{A}\boldsymbol{S}\mathrm{U} \,|\, \mathrm{C}\boldsymbol{S}\mathrm{G} \,|\, \mathrm{G}\boldsymbol{S}\mathrm{C} \,|\, \mathrm{U}\boldsymbol{S}\mathrm{A} & \text{caractères extrêmes appariés} \\
\boldsymbol{S} & \to & \boldsymbol{S}\boldsymbol{S} & \text{bifurcation} \\
\boldsymbol{S} & \to & \varepsilon & \text{terminaison}
\end{array}
$$

Cette grammaire permet d'engendrer toutes les séquences d'ARN possibles. De plus, elle est ambiguë à dessein : pour une séquence donnée, chaque arbre de dérivation possible correspond à une structure secondaire de cette séquence.

Lorsque l'on sait quels types d'appariements sont plus fréquents, on peut associer des probabilités à chaque règle. Pour une séquence donnée, il est alors possible de chercher la ou les structures secondaires les plus probables.

Qui plus est, toutes les structures secondaires ne sont pas nécessairement réalistes, et certaines formes sont plus fréquentes que d'autres. On rencontre en particulier fréquemment des imbrications d'appariements qui font apparaître des sortes de "tiges" dans les structures secondaires. Plus formellement, on appelle *tiges-boucle* un type particulier de structure secondaire sur un mot w de longueur n, pour laquelle il existe $k > 0$ tels que ses appariements sont exactement les couples $(1, n)$, $(2, n-1)$, ..., $(k, n-k)$. Les lettres qui ne sont pas appariées forment la *boucle* de cette tige-boucle.

Par exemple, la figure ci-dessous présente une tige-boucle ayant une tige de longueur 6 et une boucle de longueur 5. (Les nombres 1 et 17 indiquent le début et la fin de la séquence.)



Il est possible de définir des grammaires qui génèrent un type de structure secondaire particulier. Par exemple, voici une grammaire $\mathcal{G}_2$ qui génère toutes les tiges-boucles à trois appariements et dont la boucle est soit GCAA soit GAAA :

$$
\begin{array}{rcl}
\boldsymbol{S} & \to & \mathrm{A}\boldsymbol{W}_1\mathrm{U} \,|\, \mathrm{C}\boldsymbol{W}_1\mathrm{G} \,|\, \mathrm{G}\boldsymbol{W}_1\mathrm{C} \,|\, \mathrm{U}\boldsymbol{W}_1\mathrm{A} \\
\boldsymbol{W}_1 & \to & \mathrm{A}\boldsymbol{W}_2\mathrm{U} \,|\, \mathrm{C}\boldsymbol{W}_2\mathrm{G} \,|\, \mathrm{G}\boldsymbol{W}_2\mathrm{C} \,|\, \mathrm{U}\boldsymbol{W}_2\mathrm{A} \\
\boldsymbol{W}_2 & \to & \mathrm{A}\boldsymbol{W}_3\mathrm{U} \,|\, \mathrm{C}\boldsymbol{W}_3\mathrm{G} \,|\, \mathrm{G}\boldsymbol{W}_3\mathrm{C} \,|\, \mathrm{U}\boldsymbol{W}_3\mathrm{A} \\
\boldsymbol{W}_3 & \to & \mathrm{GAAA} \,|\, \mathrm{GCAA}
\end{array}
$$

3 Dénombrement des structures secondaires

L'ensemble des appariements associés à une structure secondaire (*sans tenir compte des lettres*) est appelé un *repliement*. On note R_n le nombre de repliements sur une

séquence de taille n.

Exemple 1 *On énumère l'ensemble des repliements d'un mot $w = w_1w_2w_3w_4$ de taille 4.*

Dans la suite, quand on donne un mot, on note en gras les lettres appariées – noter que cela ne suffit pas à décrire de manière unique un repliement ; par exemple, dans le mot $\boldsymbol{w_1w_2w_3w_4}$, w_1 peut être apparié soit avec w_2, soit avec w_4.

Soit

- w_1 n'est pas apparié. *On obtient les appariements correspondants comme w_1 concaténé avec un repliement de $w_2w_3w_4$, ce qui donne*
$$w_1 \cdot \boldsymbol{w_2}w_3\boldsymbol{w_4};\ w_1 \cdot \boldsymbol{w_2w_3}w_4;\ w_1 \cdot w_2\boldsymbol{w_3w_4};\ w_1 \cdot w_2w_3w_4.$$
- w_1 est apparié. *Il y a alors 3 possibilités :*
  - *w_1 est apparié avec w_2 ; on obtient alors la concaténation de $\boldsymbol{w_1w_2}$ avec un appariement de w_3w_4, soit les deux possibilités*
    - $\boldsymbol{w_1w_2} \cdot w_3w_4$;
    - $\boldsymbol{w_1w_2} \cdot \boldsymbol{w_3w_4}$;
  - *w_1 est apparié avec w_4 ; on obtient alors de même les possibilités :*
    - $\boldsymbol{w_1} \cdot w_2w_3 \cdot \boldsymbol{w_4}$;
    - $\boldsymbol{w_1} \cdot \boldsymbol{w_2w_3} \cdot \boldsymbol{w_4}$;
  - *enfin, si w_1 est apparié avec w_3 on obtient $\boldsymbol{w_1}w_2\boldsymbol{w_3}w_4$.*

On a $R_1 = 1$, $R_2 = 2$, $R_3 = 4$, $R_4 = 9$ et on observe que

$$R_4 = R_3 + 2R_2 + R_1^2\ . \tag{5}$$

Faisons la convention que $R_0 = 1$. Plus généralement, on obtient la relation de récurrence,

Théorème 1 *La suite $(R_n)_{n\geq 0}$ vérifie la relation de récurrence*

$$R_0 = R_1 = 1, R_n = R_{n-1} + \sum_{k=0}^{n-2} R_kR_{n-2-k} \quad \forall n \geq 2. \tag{6}$$

Cette récurrence peut être utilisée pour étudier expérimentalement l'asymptotique de R_n et observer que R_n croît exponentiellement ; une étude mathématique (qu'on ne cherchera pas à reproduire) permet plus précisément d'obtenir l'approximation, valable pour n grand,

$$R_n \approx \sqrt{\frac{27}{4\pi n}}\frac{3^n}{n}.$$

Réflechissons un peu à ce que ce résultat sur les repliements implique sur les structures secondaires. Pour un repliement de longueur n donné, le nombre de séquences d'ARN pouvant avoir ce repliement comme structure secondaire est 4^{n-t}, où $t < n/2$ est le nombre d'appariements du repliement.

Par suite, le nombre total de structures secondaires de longueur n est au moins de l'ordre de 6^n, et le nombre moyen par séquence ADN est au moins de l'ordre de $1{,}5^n$, donc exponentiel comme annoncé. L'énumération des structures secondaires pour trouver la plus probable n'est donc effectivement pas envisageable.

4 Pistes de réflexion pour l'exposé

1. Dessiner la ou les structures secondaires optimales correspondant à l'exemple de la Table 1.

2. Dans les structures d'ARN réelles, deux caractères consécutifs ne peuvent pas être appariés. Modifier les équations $(1) - (3)$ pour que cette contrainte soit prise en compte. Comment modifier la première grammaire ?

3. Compléter l'algorithme pour que, en fonction de la matrice résultat, il affiche une structure secondaire optimale. Cette structure optimale est-elle nécessairement unique ?

4. Évaluer la complexité de cet algorithme, en espace et en temps.

5. Écrire l'algorithme et le programme correspondant à la partie 2.2.2.

6. Compléter la réponse à la piste précédente en écrivant une fonction qui prend en entrée deux séquences w et w' et détecte si w et w' peuvent être appariées en au moins une position, ie. s'il existe un i tel que w_i et w'_i soient complémentaires.

7. Écrire une grammaire qui engendre toutes les tiges-boucles dont la longueur des boucles est supérieure ou égale à 3.

8. La grammaire $\mathcal{G}_1$ est volontairement ambiguë, afin que l'ensemble des structures secondaires possibles pour chaque séquence puisse être engendré. Cependant, avec cette grammaire, pour une même séquence plusieurs arbres de dérivation peuvent donner la même structure secondaire. En donner un exemple, puis proposer une autre grammaire avec laquelle chaque structure secondaire donne lieu à un unique arbre de dérivation.

9. Valider expérimentalement l'approximation finale ; on pourra également estimer l'ordre de grandeur de l'erreur d'approximation.

10. La prédiction de structures d'ARN s'effectue sur des séquences issues des techniques de séquençage à haut débit de génomes entiers. Le génome humain, par exemple, contient de l'ordre de 3 milliards de paires de bases. Discuter au choix d'un des points suivants : sécurité de la conservation de ces données, utilisation de données personnelles sensibles, sobriété numérique.