bien sûr appuyé sur des notions bien connues dans la littérature comme la recherche par fenêtre glissante, ou l'idée de la LookAheadTable qui est déjà mentionné dans l'article de Beckstette et al. 2006 comme faisant partie du « folklore »).

Dans le reste de ce mémoire, j'utilise le pronom nominatif « nous » plutôt que « je », car cela reflête mieux la réalité du travail de recherche.

2 Définition du problème et notations

Avant de définir formellement notre problème, nous déterminons les notations utilisées dans ce rapport.

2.1 Notations

Il est important de préciser que nous présentons dans ce rapport un alphabet quelconque, mais dans la pratique, nous travaillons sur de l'ADN et notre alphabet est de taille 4.

— Pour un alphabet Σ de taille σ, une di-PWM P représentant un motif de taille m est une matrice de taille $\sigma^2 \times (m-1)$. Le poids du dinucléotide db où d et b appartiennent à Σ aux positions respectives i et $i+1$ sera alors noté $P[db, i]$. Il correspond au *log* de la fréquence observée du dinucléotide à la position donnée rapportée à sa fréquence attendue(1/16). La figure 1 est une représentation d'un motif de taille 6 et ayant pour alphabet A, C, G, T.

	0	1	2	3	4
AA	0.77	-0.14	-0.14	-1.97	-1.77
AC	-0.78	-1.97	-2.58	-2.58	-2.23
AG	-0.65	0.50	1.08	-4.4	-1.45
AT	-1.77	-1.2	-4.4	1.27	-4.4
CA	0.52	0.26	-1.11	-3.18	0.64
CC	-1.77	-0.43	-3.12	-0.56	-1.6
CG	-1.32	0.94	0.31	-4.4	-0.14
CT	-3.12	-0.22	-4.4	-1.77	1.97
GA	1.48	0.16	0.82	-1.45	-3.12
GC	0.33	-0.43	-2.23	-1.97	-4.4
GG	1.28	0.85	1.83	-1.97	-1.97
GT	-1.02	-1.32	-2.23	2.42	0.16
TA	-1.21	-0.59	-1.11	-4.4	0.61
TC	-2.23	-1.11	-4.4	-4.4	2.14
TG	-1.45	0.78	0.19	-4.4	-0.54
TT	-2.23	-1.45	-4.4	-2.23	0.16

di-PWM

FIGURE 1 – Di-PWM représentant un motif de taille 6 et ayant pour alphabet A, T, G et C. Les colonnes correspondent aux positions dans le motif et les lignes aux dinucléotides. Une valeur représente le poids d'un dinucléotide à une position donnée. Cette di-PWM est une di-PWM factice inspirée de la di-PWM du FT ATF3 réduite en longueur et dont les positions ont été modifiées.

— Le seuil de score fixé comme paramètre de la recherche est noté θ.

— Le texte dans lequel on effectue la recherche est noté T et sa longueur n.

— Pour un mot $u = u_0 \dots u_l$ et $0 \leq i < j \leq l$, on note $u[i, j]$ la sous-chaîne de u commençant à la position i et se terminant à la position j.

— Pour un mot de longueur m, $u = u_0 \dots u_{m-1}$, un préfixe de u est une sous-chaîne de u ayant pour premier élément u_0 et de longueur strictement inférieur à m. Un suffixe de u est une sous-chaîne de u commençant à une position strictement supérieure à 0 et ayant pour dernière position u_{m-1}.

— On note Σ^p où $p \in \mathbb{N}$ l'ensemble des mots possibles de longueur p formés sur l'ensemble de l'alphabet Σ.

Avec un motif probabiliste de type PWM ou di-PWM, pour un mot u de longueur m on peut calculer le score de u selon P. Le score est déterminé par la somme des valeurs pour chaque dinucléotide de u en fonction de leur position.

Calcul du score Le score d'un mot u de taille m pour une di-PWM P de taille $m-1$, noté $score(u)$ est calculé de la façon suivante :

$$score(u) = \sum_{i \in \{0, m-2\}} P[\ u[i, i+1], i]$$

Le calcul de ce score est une somme et celle-ci étant commutative, il peut se faire dans un ordre différent. On peut le calculer indifféremment de gauche à droite ou de droite à gauche, ou en changeant l'ordre des positions.

Calcul du seuil θ en fonction du ratio seuil donné On note $score_{max}$ et $score_{min}$ les scores maximum et minimum d'une di-PWM et $ratio$ le ratio du seuil donné en entrée.

$$\theta = (score_{max} - score_{min}) * ratio + score_{min}$$

— On définit une occurrence comme une sous-chaîne de T dont le score est supérieur ou égal à θ.

Notre problème se définit de la manière suivante. Nous cherchons la position et le score des occurrences du motif P dans le texte T.

2.2 Calcul des majorants

Nous utilisons dans nos différents algorithmes une estimation d'un score maximum atteignable pour un mot de taille m. Ce score est calculé à partir du score du préfixe du mot qu'on nomme score partiel auquel on ajoute un majorant du suffixe de ce mot. Afin de calculer le score maximum atteignable pour un mot, nous pouvons utiliser l'une des deux structures distinctes que sont la *LookAheadTable(LAT)* et la *LookAheadMatrix(LAM)*. La figure 2 présente ce principe de manière générale.

Les deux structures se construisent de manière semblable. Les deux algorithmes de construction commencent par calculer le(s) dernier(s) élément(s) de la structure

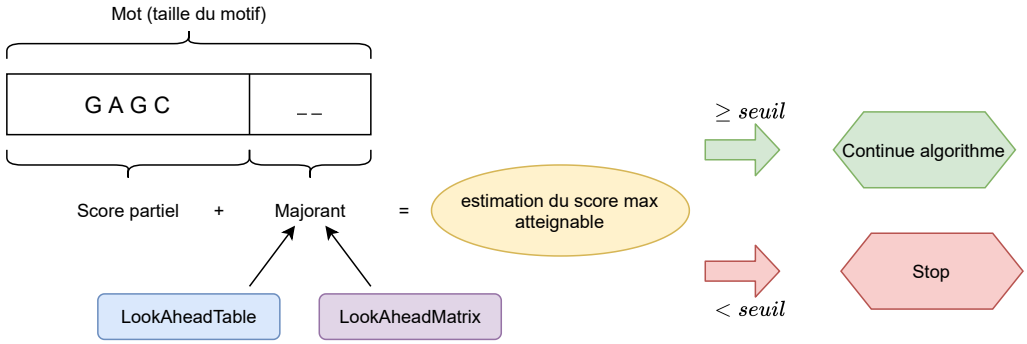

FIGURE 2 – Estimation du score maximum atteignable pour un mot de taille m à partir du score partiel de son préfixe. Le score du préfixe est calculé (score partiel) et sommé au majorant calculé à partir de la LAT ou de la LAM. Le score maximum atteignable pour un mot complet à partir de ce préfixe est évalué en comparaison au seuil fixé afin de déterminer si l'algorithme doit continuer ou non.

en partant de la fin du motif. Puis, ils avancent vers le début du motif en calculant les éléments de manière récursive. La dernière position est traitée indépendemment, par la suite, les positions suivantes dans l'ordre du traitement, dépendent du calcul précédent.

Nous détaillons ensuite la définition et l'algorithme de construction de la LAT puis ceux de la LAM.

2.2.1 La *LookAheadTable*

La LAT d'une di-PWM est une table qui va stocker le majorant que le score d'un suffixe peut atteindre à partir d'une position donnée. La définition se trouve en définition 2.1.

Definition 2.1 (*LookAheadTable* associée à une di-PWM)**.**
La LAT L d'une di-PWM P d'un motif de taille m est une table de taille $(m-1)$ où pour tout entier i tel que $0 \leq i \leq m-2$

$$L[i] := \begin{cases} \max\limits_{d,b \in \Sigma} P[db, i], & \text{si } i = m-2 \\ \max\limits_{d,b \in \Sigma} P[db, i] + L[i+1], & \text{si } 0 \leq i < m-2 \end{cases}$$

Ici db représente le dinucléotide constitué de la paire de symboles d et b.

$L[i]$ est un majorant du score d'un suffixe de longueur $(m-1-i)$. L'algorithme 1 calcule la LAT d'un di-PWM. Comme la calcul de $L[i]$ dépend de $L[i+1]$, le calcul de la LAT se fait de droite à gauche dans la di-PWM (cf. la boucle ligne 8 algorithme 1).

2.2.2 La *LookAheadMatrix*

La LAM d'une di-PWM est une matrice qui va stocker le score maximum qu'un suffixe peut atteindre à partir d'une position donnée et commençant par une lettre de l'alphabet spécifique, cf. définition 2.2.