TD / TME
Recherche de Motifs

Définir un **motif commun** à un ensemble de séquences revient à construire un **profil** qui soit le plus représentatif des séquences considérées. Le profil est une **expression régulière** écrite selon les conventions symboliques IUPAC :

R	A + G	pu**R**ines
Y	T + C	p**Y**rimidines
M	A + C	groupe a**M**ino
K	G + T	groupe **K**eto (cétone)
W	A + T	**W**eak (faible)
S	G + C	**S**trong (forte)
B	G + C + T	**Not A**
D	A + G + T	**Not C**
H	A + T + C	**Not G**
V	A + G + C	**Not T**
N	A + G + C + T	any **N**ucleotide

<G	Le nucléotide G est au début de la séquence du motif
A	Le nucléotide A est à la position donnée du motif
X	N'importe quel nucléotide est toléré à cette position du motif
[AC]	Liste qui représente la possibilité d'avoir un des nucléotides cités à la position donnée. Seuls A ou C sont possibles à cette position mais jamais T ou G. La dégénérescence concerne cette position.
{T}	Liste d'exclusion : le nucléotide T ne doit jamais être retrouvé à cette position. A, C ou G sont possibles.
[CT] (2)	L'entier entre parenthèses indique un nombre de répétitions consécutives. A cette position, on pourra trouver soit CC soit TT.
T (1,2)	Le nucléotide T est répété entre une et deux fois
-	Symbole ou élément qui sépare les résidus du motif
A>	Le nucléotide A est à la fin de la séquence du motif

EXERCICE 1 :

On donne le motif suivant :

[KR] − x(1,3) − [RKSAQ] − N − {VL} − x − [SAQ](2) − {L} − [RKTAENQ] − x − R − {S} − [RK]

1. Quelle est la taille de ce motif ?
2. Interpréter les différentes positions de ce profil protéique.
3. Quelle est la signification de x (1,3) ?

EXERCICE 2 :

On donne le résultat de l'alignement multiple suivant :

-3	-2	-1	1	2	3	4	5	6
C	G	G	G	T	A	A	G	T
A	A	G	G	T	A	T	G	C
C	A	G	G	T	G	A	G	G
T	G	G	G	T	A	A	C	T
C	A	A	G	T	A	A	G	C
A	A	G	G	T	A	G	G	C
A	T	G	G	T	G	A	G	T
T	T	G	G	T	A	A	G	G
A	A	G	G	T	A	T	T	T
A	A	G	G	T	A	A	G	G

1. Donner la séquence consensus ou motif en utilisant le code IUPAC
2. Calculer la table des fréquences, en déduire la matrice des poids-position
3. Déterminer la structure finale du motif et calculer son score

EXERCICE 3 :

On considère les séquences suivantes :

A A A A C T G T G G
A A A A C T G T G G
C A A A T T G T G G
A A A A A T G T C G
A A A A C C A T G C

1. Donner la séquence consensus ou motif en utilisant le code IUPAC
2. Calculer la table des fréquences, en déduire la matrice des poids-position
3. Déterminer la structure finale du motif et calculer son score
4. Soit la séquen**e AAAGCTGTCC**. Calculer son score et conclure.

EXERCICE 4 :

On considère les séquences suivantes :

ACCACTACTCATCATCATTCATCGGCTGTGC
GACTCGATCAGCATTATHGGGCTGTCGTAG
GGCTGGCTGTACTATTATTAGCTAGCTGATG
GGCTGCTGACAGGATGATAGCTACGATCGA

1. Procéder à l'alignement multiple de ces séquences
http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::clustalw-multialign
2. Utiliser l'alignement obtenu pour identifier la région de plus haute conservation et construire un profil de cette région
3. Calculer la matrice poids-position à partir de cet alignement
4. Prendre une séquence quelconque et calculer la probabilité qu'elle soit une occurrence du motif.
5. Rechercher les motifs présents dans ces séquences avec l'outil en ligne MEME : http://meme.nbcr.net/meme/cgi-bin/meme.cgi. Discuter les résultats.

EXERCICE 5 :

La méthode **MEME** (**M**ultiple **E**xpectation-**M**aximisation **M**otif **E**licitation) permet de détecter des motifs dans un ensemble de séquences ADN ou protéiques reliées, non alignées. En particulier, étant donné un groupe de séquences de longueur *L*, dont on sait qu'elles partagent un motif commun de longueur *W,* l'algorithme **MEME (OOS)** :

- Infère un modèle (Θ, p_0, Z) pour le motif
- Localise l'occurrence du motif dans chaque séquence

Θ est la matrice des poids-position $p_{c,k}$ du motif, avec c $\in$ {A,C,G,T} et k $\in$ {1…*W*}, p_0 est le vecteur de probabilités du modèle nul. Z est la matrice des variables cachées, qui donnent les positions initiales du motif : $Z_{i,j} = 1$ si le motif commence en position *j* de la séquence *i*, $Z_{i,j} = 0$ sinon. L'algorithme affine les paramètres du modèle de manière itérative par espérance-maximisation. Chaque itération *t* se compose de deux étapes :

- **(E)** Calcul des valeurs attendues $Z^{(t)}$ de Z, étant donné $\Theta^{(t-1)}$, $p_0^{(t-1)}$

$$\mathrm{Z}_{i,j}^{(t)} = P(\mathrm{Z}_{i,j} = 1 | X_i, \Theta^{(t-1)}, p_0^{(t-1)}) = \frac{P(X_i | \mathrm{Z}_{i,j} = 1, \Theta^{(t-1)}, p_0^{(t-1)})}{\sum_{k=1}^{L-W+1} P(X_i | \mathrm{Z}_{i,k} = 1, \Theta^{(t-1)}, p_0^{(t-1)})} \quad \textit{(Formule de Bayes)}$$

où

$$P(X_i | \mathrm{Z}_{i,j} = 1, \Theta^{(t-1)}, p_0^{(t-1)}) = \prod_{k=1}^{j-1} p_{c_k,0} \prod_{k=j}^{j+W-1} p_{c_k,k-j+1} \prod_{k=j+W}^{L} p_{c_k,0}$$

- **(M)** Estimation des paramètres du modèle, avec « pseudo-counts » *d*

$$p_{c,k}^{(t)} = \frac{n_{c,k} + d_{c,k}}{\sum_b n_{b,k} + d_{b,k}} \quad \textit{où} \quad n_{c,k} = \sum_i \sum_{\{j | X_{i,j+k-1} = c\}} \mathrm{Z}_{i,j}^{(t)} \quad \textit{si k>0 (motif) et} \quad n_{c,0} = n_c - \sum_{j=1}^{W} n_{c,j} \quad \textit{(modèle nul)}$$

Ce processus est répété jusqu'à ce que les paramètres ou la vraisemblance du modèle n'évoluent plus. La vraisemblance totale du modèle est exprimée comme suit :

$$P(D | \Theta, p_0) = \prod_i P(X_i | \Theta, p_0) = (L - W + 1)^{-n} \prod_i \sum_j P(X_i | \mathrm{Z}_{i,j} = 1, \Theta, p_0)$$

Pour initialiser la matrice poids-position, on prend généralement un motif au hasard dans les séquences, et on fixe à 0.5 les poids correspondant au motif et (1-0.5)/3 les autres. Le modèle nul par défaut est p_0=(0.25, 0.25, 0.25, 0.25).

Pour une illustration : http://www.biostat.wisc.edu/bmi776/lectures/motif-modeling.pdf

1. Ecrire un programme qui prend en entrée un ensemble de séquences ADN et une largeur *W*, et apprend un modèle de motif pour un motif de largeur *W*. Le programme devra renvoyer les paramètres du modèle (Θ, p_0) et la liste des meilleures positions du motif dans chaque séquence (Z). Vous calculerez la valeur de la log-vraisemblance totale du modèle $\log P(D | \Theta, p_0)$ à chaque itération et l'algorithme prendra fin lorsque $\Delta \log P(D | \Theta, p_0) < \varepsilon << 1$. Vous partirez de

conditions initiales multiples.

2. Identifier le mot commun de largeur 14 dans les séquences du fichier : http://www.biostat.wisc.edu/bmi776/hw/hw1_hidden_motif.txt
3. Construire une séquence LOGO pour le motif prédit avec le service *WebLogo*.
4. Rechercher la base de données JASPAR (http://jaspar.cgb.ki.se/) en utilisant la matrice de poids-position. Quel facteur de transcription se fixe sur ce motif?