M1 Bioinformatique, Connaissances et Données
Master Sciences et Numérique pour la Santé — Année 2016-2017
HMSN206 - Partie Alignement

TD 2 : Consensus, Motif, Profil

Anne-Muriel Arigon Chifolleau - Université Montpellier, LIRMM

chifolleau@lirmm.fr

1 Consensus

Code	Description
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A
K	T, U, or G
W	T, U, or A
S	C or G
B	C, T, U, or G (not A)
D	A, T, U, or G (not C)
H	A, T, U, or C (not G)
V	A, C, or G (not T, not U)
N	Any base (A, C, G, T, or U)

Voici ci-contre, le code IUPAC (International Union of Pure and Applied Chemistry) pour les nucléotides.

Une séquence **consensus** est une séquence pour laquelle à chaque position de l'alignement multiple, on retient la lettre majoritaire.

Ci-dessous, un alignement multiple du site de fixation du facteur de transcription *c-Ets-1* chez les 15 murins (source TRANSFAC M00032).

Trouver le consensus pour ce site de fixation ainsi que son code IUPAC.

	G	C	C	G	G	A	A	G	T	G
	A	C	C	G	G	A	A	G	C	A
	G	C	C	G	G	A	T	G	T	A
	A	C	C	G	G	A	A	G	C	T
	A	C	C	G	G	A	T	A	T	A
	C	C	C	G	G	A	A	G	T	G
	A	C	A	G	G	A	A	G	T	C
	G	C	C	G	G	A	T	G	C	A
	T	C	C	G	G	A	A	G	T	A
	A	C	A	G	G	A	A	G	C	G
	A	C	A	G	G	A	T	A	T	G
	T	C	C	G	G	A	A	A	C	C
	A	C	A	G	G	A	T	A	T	C
	C	A	A	G	G	A	C	G	A	C
	T	C	T	G	G	A	C	C	C	T
Séquence consensus :										
Code IUPAC :										

2 Représentation d'un motif par une matrice

2.1 Comptage

Dans la matrice que nous voulons remplir, les lignes représentent les colonnes de l'alignement multiple et les colonnes de la matrice, les acides nucléiques (4 colonnes).

Avec l'exemple de l'exercice précédent, remplissez la matrice ci-dessous. La première ligne est donnée à titre d'exemple.

A	C	G	T
7	2	3	3

2.2 Matrice de fréquence et matrice poids position

Une fois le comptage effectué, on détermine la fréquence de chaque nucléotides dans chaque colonne du MSA (valeur entre 0 et 1). Pour passer à une matrice poids-position il faut encore une opération supplémentaire. L'idée est que les bases qui apparaissent plus que la moyenne aient un score positif et celles qui apparaissent moins que la moyenne, un score négatif. On applique donc la formule suivante où $f(x)$ représente la fréquence de x.

$$\text{Poids de la base } x \text{ dans une colonne de l'alignement : } \log \frac{f(x)}{0.25}$$

Que représente 0.25 ?

Les matrices de fréquences et poids-position correspondant à notre exemple sont données ci-dessous.

A	C	G	T
0.47	0.13	0.2	0.2
0.07	0.93	0	0
0.33	0.6	0	0.07
0	0	1	0
0	0	1	0
1	0	0	0
0.53	0.13	0	0.33
0.27	0.07	0.67	0
0.07	0.4	0	0.53
0.33	0.27	0.27	0.13

Matrice de fréquences

A	C	G	T
0.91	-0.94	-0.32	-0.32
-1.8	1.9	-2.3	-2.3
0.4	1.26	-2.3	-1.8
-2.3	-2.3	2	-2.3
-2.3	-2.3	2	-2.3
2	-2.3	-2.3	-2.3
1.1	-0.94	-2.3	0.4
0.11	-1.8	1.42	-2.3
-1.8	0.68	-2.3	1.1
0.4	0.11	0.11	-0.94

Matrice poids-position

2.3 Calcul du score d'un mot

Une fois la matrice poids-position calculée. On s'intéresse au score d'un mot dans ce modèle. Pour calculer le score d'un mot, il suffit d'additionner les poids correspondant aux lettres qui le compose, position par position. Considérons la matrice poids-position de la feuille précédente.

1. Comment sait-on qu'un score sera bon ou mauvais ?
2. Calculer le meilleur score possible avec cette matrice ; quel motif atteint ce score ?
3. Le moins bon score ; de même, quel motif atteint ce score ?
4. Calculer le score du mot TACGGATACG

3 Motif PROSITE

Nous allons maintenant créer un motif PROSITE correspondant à l'alignement multiple suivant d'hormone pancréatique. (Source : PROSITE, PS00265)

NEUY CARAU/29-64	A	E	E	.	.	L	A	K	Y	Y	S	A	L	R	H	Y	I	N	L	I	T	R	Q	R	Y
PYY HUMAN/29-64	P	E	E	.	.	L	N	R	Y	Y	A	S	L	R	H	Y	L	N	L	V	T	R	Q	R	Y
PMY PETMA/1-36	P	E	E	.	.	L	S	K	Y	M	L	A	V	R	N	Y	I	N	L	I	T	R	Q	R	Y
PPY LOPAM/1-36	P	E	D	.	.	W	A	S	Y	Q	A	A	V	R	H	Y	V	N	L	I	T	R	Q	R	Y
PAHO BOVIN/30-65	P	E	Q	.	.	M	A	Q	Y	A	A	E	L	R	R	Y	I	N	M	L	T	R	P	R	Y
PAHO CHICK/26-61	V	E	D	.	.	L	I	R	F	Y	N	D	L	Q	Q	Y	L	N	V	V	T	R	H	R	Y
PAHO ANSAN/1-36	V	E	D	.	.	L	R	F	Y	Y	D	N	L	Q	Q	Y	R	L	N	V	F	R	H	R	Y
NPF HELAS/4-39	P	N	E	.	.	L	R	Q	Y	L	K	E	L	N	E	Y	Y	A	I	M	G	R	T	R	F
NPF MONEX/1-39	D	N	K	A	A	L	R	D	Y	L	R	Q	I	N	E	Y	F	A	I	I	G	R	P	R	F
									*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*

La syntaxe pour la description du motif est la suivante :
— Tous les éléments de l'expression sont séparés par des tirets **-**
— Le joker est la lettre **X** pour les protéines (**N** pour l'ADN)
— On peut préciser le nombre d'occurrences avec des parenthèses : **X(5)** ou **D(2,4)**
— Le choix entre plusieurs acides aminés possibles se note avec des crochets **[AP]**

1. À partir de l'alignement multiple ci-dessous, construisez un motif pour le site actif présumé.
2. Combien de "chaînes de caractères" correspondent au motif que vous venez de trouver ?
3. Combien de "chaînes de caractères" correspondent au motif `'D-E-X(3)-[LIVM]-X(1,3)-[FY]'` ?

4 Modèle de Markov caché (HMM)

Vous devez construire les profils HMM des 2 alignements multiples ci-dessous .

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

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