2. Motifs nucléiques et matrice de poids

L'écriture de l'expression régulière d'une séquence consensus en se basant sur le code IUPAC:

position / sequences	1	2	3	4	5	6	7	8	9	10
seq 1	G	C	C	G	G	A	A	G	T	G
seq 2	A	C	C	G	G	A	A	G	C	A
seq 3	G	C	C	G	G	A	T	G	T	A
seq 4	A	C	C	G	G	A	A	G	C	T
seq 5	A	C	C	G	G	A	T	A	T	A
seq 6	C	C	C	G	G	A	A	G	T	G
seq7	A	C	A	G	G	A	A	G	T	C
seq 8	G	C	C	G	G	A	T	G	C	A
seq 9	T	C	C	G	G	A	A	G	T	A
seq 10	A	C	A	G	G	A	A	G	C	G
seq 11	A	C	A	G	G	A	T	A	T	G
seq 12	T	C	C	G	G	A	A	A	C	C
seq 13	A	C	A	G	G	A	T	A	T	C
seq 14	C	A	A	G	G	A	C	G	A	C
seq 15	T	C	T	G	G	A	C	C	C	T
sequence consensus selon le code IUPAC	N	C	H	G	G	A	H	V	H	N

Figure 2 : l'exemple suivant représente l'expression d'une séquence consensus résultante d'un alignement multiple selon le code IUPAC.

Il existe une autre représentation graphique de ces motifs. Le principe des logos dans cette présentation stipule que la taille de la base est proportionnelle à sa fréquence sur la position de l'alignement. L'exemple au dessous est réalisé avec le programme WEBLOGO fig.3.

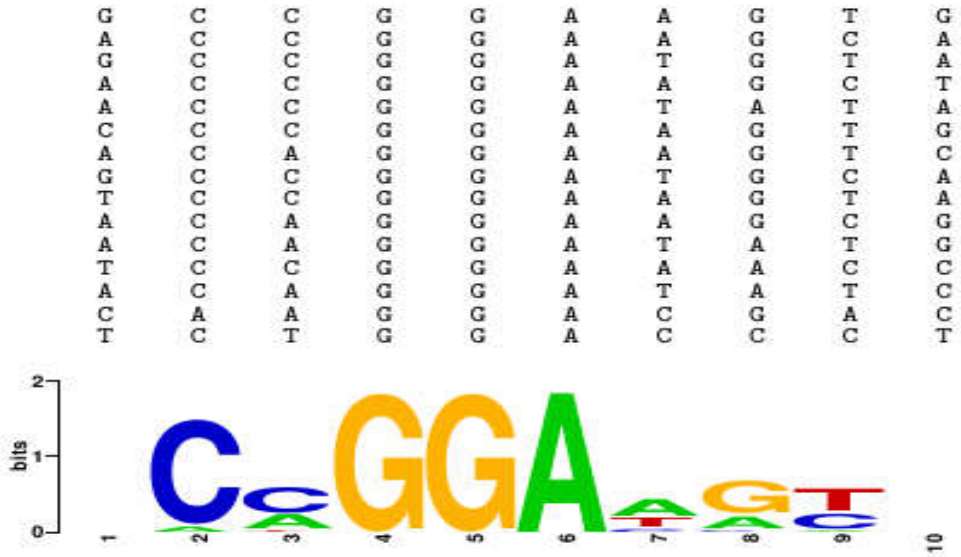

Figure3 : représentation graphique d'une séquence consensus résultante d'un alignement multiple par le programme WEBLOGO.

On utilise une représentation ou une écriture de la séquence consensus de manière plus précise par rapport au code IUPAC, en appliquant les expressions régulières « les conventions symboliques » ou la grammaire PROSITE présentées dans le tableau au dessous :

Tableau 2 : les conventions symboliques pour l'écriture d'une séquence consensus :

T	Une seule lettre représente la base biologique à la position donnée. Dans cet exemple la thymine est représentée par T
[ATC]	Une liste représente la possibilité de trouver une des bases composent la liste pour une position. Soit A, T ou C dans cet exemple
{ C }	Liste d'exclusion : pas de C (cytosine) dans cette position
N(pour AN) ou X (pour protéines)	N'importe quelle base biologique
{AC} (2)	Soit deux A ou deux C
X (0 .1)	Entre 0 et 1 base quelconque
< C	La séquence est commencée par C
T >	La séquence est terminée par T

Exemple :
<[AT]-G-x(3)-A-[ATG] -T>, ce motifs signifie qu'en première position il y a un A ou un T, puis un G puis 3 bases inconnues puis A , puis l'une de trois bases soit A ou T ou G , T en dernière position.

La matrice de pondération : PWMs : Position Weight Matrice ou PSSMs : Position specific scoring Matrice, ou profile:
Pour exprimer l'ambiguïté et la complexité d'un motif, on peut également déduire de l'alignement des séquences une table de fréquences en comptabilisant les occurrences(les apparitions) de chaque base à chaque position du motif.

Exemple: soit l'alignement multiple suivant :

position / séquence	1	2	3	4	5	6
Seq 1	A	T	T	C	T	G
Seq 2	C	T	G	G	G	G
Seq 3	T	T	G	C	T	A
Seq 4	T	T	A	C	T	G
Seq5	T	T	G	C	T	A
Seq 6	A	T	G	C	T	A
Séquence consensus Selon IUPAC	**W**	**T**	**K**	**C**	**T**	**R**

Nous construisons une table de fréquence pour déterminer la probabilité d'apparition d'une base dans chaque position du motif.

Base / Position	A	T	C	G	Total
1	2	3	1	0	6
2	0	6	0	0	6
3	1	1	0	4	6
4	0	0	5	1	6
5	0	5	0	1	6
6	3	0	0	3	6
Total	6	15	6	9	36

Nous calculons la fréquence ***fi*** rapportée au nombre totale des séquences

$$fi = \frac{\text{nbre d'occurance de chaque base dans chaque position}}{\text{nbre total de sequence}}$$

La fréquence ***Fi*** on utilisant la formule suivante :

$$Fi = \frac{\text{nbre total d'occurance de chaque base}}{\text{nbre total de base}}$$

Base / Position	A	T	C	G
1	2/6	3/6	1/6	0
2	0	6/6	0	0
3	1/6	1/6	0	4/6
4	0	0	5/6	1/6
5	0	5/6	0	1/6
6	3/6	0	0	3/6
Fi	6/36	15/36	6/36	9/36

Matrice de poids : pour calculer le poids de chaque base: on utilise le rapport ***fi/Fi***

Base / Position	A	T	C	G
1	1.98	1.21	1.00	0.00
2	0.00	2.43	0.00	0.00
3	1.00	0.39	0.00	2.64
4	0.00	0.00	2.03	0.64
5	0.00	2.03	0.00	0.64
6	3.12	0.00	0.00	2.00

Nous calculons le Log (*fi/Fi)* de chaque ligne de la matrice de poids et on obtient la matrice suivante :

	A	T	C	G
1	0.29	0.08	0.00	-10
2	-10	0.38	-10	-10
3	0.00	-0.40	-10	0.42
4	-10	-10	0.30	-0.19
5	-10	0.30	-10	-0.19
6	0.49	-10	-10	0.30

A partir de cette matrice, on écrit l'expression exacte de la séquence consensus:

1	2	3	4	5	6
A_T	**T**	**G**	**C**	**T**	**A_G**