Les motifs

Introduction

Le motif (ou encore signature) peut être défini comme étant une courte séquence (un segment court) continue, non ambigüe et peu dégénérée. C'est une zone fortement conservée le long de l'évolution qui est composée de quelques résidus et est commune à un ensemble de séquences (nucléiques ou protéiques) ayant la même fonction et le même mécanisme biologiques (séquences homologues). Schématiquement, on peut imaginer le motif comme un petit dessin qui se répète ou non le long d'une séquence donnée. Ce petit dessin sera presque identique chez un bon nombre de séquences ; mais jamais dans les mêmes positions à l'intérieur de ces différentes séquences le contenant :

Le terme MOTIF est remplacé par le mot ***pattern*** chez les anglo-saxons, sauf que celui-ci peut contenir plusieurs motifs à la fois : Le pattern est une séquence dégénérée et/ou composée de différents motifs avec régions variables. Le motif peut être impliqué dans des fonctions biologiques ou dans des systèmes de régulations. Comme il peut servir également à identifier une séquence inconnue après confrontation à une base de motifs.

Actuellement, on peut aisément consulter deux bases de motifs nucléiques spécifiques et qui sont régulièrement mises à jour : ce sont les bases de facteurs de transcription **TFD** (Ghosh, 1993), et **TRANSFAC** (Knüppel et al., 1994). Par exemple, la base TFD correspond aux facteurs de transcription des eucaryotes. Elle contient des informations relatives aux motifs mais également des informations qui concernent les séquences protéiques concernées par la transcription comme par exemple les domaines protéiques en liaison avec l'ADN.

Définition des consensus

Définir un motif commun à un ensemble de séquences, c'est l'identifier et l'écrire convenablement de façon conventionnelle (Code IUPAC) pour qu'il soit le plus "*représentatif*" de toutes les séquences. La définition d'un motif commence le plus souvent par un alignement multiple des séquences supposées contenir le motif. Une fois les séquences alignées, la zone commune supposée le motif sera sélectionnée à part pour servir à sorte d'annotation pour confirmer s'il s'agit bien d'un motif ou non. Il existe plusieurs façons de définir un motif : On peut par exemple, confronter la zone obtenue avec une base de motifs. La méthode des matrices de poids-position va nous permettre de définir le motif.

Prenons l'exemple d'alignement suivant :

Position / Individu	1	2	3	**4**	5	**6**	7	8	9	**10**	11	**12**
1	T	T	C	**C**	C	**G**	A	A	G	**G**	C	**A**
2	C	T	T	**C**	G	**G**	G	T	T	**G**	T	**A**
3	C	T	T	**C**	G	**G**	A	T	T	**G**	T	**A**
4	T	T	T	**C**	G	**G**	A	T	C	**G**	T	**A**
5	T	T	C	**C**	C	**G**	A	A	G	**G**	G	**A**
6	T	C	C	**C**	C	**G**	A	A	G	**G**	G	**A**
7	T	C	C	**C**	C	**G**	A	A	G	**G**	T	**A**
8	T	T	C	**C**	G	**G**	A	A	G	**G**	T	**A**
9	T	T	C	**C**	G	**G**	A	T	G	**G**	T	**A**
10	C	T	C	**C**	G	**G**	G	A	T	**G**	T	**A**

La première lecture du tableau, révèle quatre positions communes à tous les individus (4 = C, 6 = G, 10 =
G et 12 = A). Ces nucléotides vont obligatoirement être retrouvés dans l'écriture finale du motif aux mêmes
positions. Le motif final sera écrit alors sous la forme :

1	2	3	**4**	5	**6**	7	8	9	**10**	11	**12**
N	N	N	**C**	N	**G**	N	N	N	**G**	N	**A**

Il nous reste donc à déterminer quels seront les nucléotides qui vont occuper le reste des positions à savoir 1, 2, 3, 5, 7, 8, 9 et 11. Pour cela, il faut passer par un nombre de calculs.

4. Table des fréquences

Dans un premier temps, on calcule le nombre d'occurrences de chaque nucléotide par colonne ainsi que les totaux:

Position / Individu	1	2	3	4	5	6	7	8	9	10	11	12	TOTAL
1	T	T	C	C	C	G	A	A	G	G	C	A	
2	C	T	T	C	G	G	G	T	T	G	T	A	
3	C	T	T	C	G	G	A	T	T	G	T	A	
4	T	T	T	C	G	G	A	T	C	G	T	A	
5	T	T	C	C	C	G	A	A	G	G	G	A	
6	T	C	C	C	C	G	A	A	G	G	G	A	
7	T	C	C	C	C	G	A	A	G	G	T	A	
8	T	T	C	C	G	G	A	A	G	G	T	A	
9	T	T	C	C	G	G	A	T	G	G	T	A	
10	C	T	C	C	G	G	G	A	T	G	T	A	
A	0	0	0	0	0	0	8	6	0	0	0	10	24
C	3	2	7	10	4	0	0	0	1	0	1	0	28
G	0	8	0	0	6	10	2	0	6	10	2	0	44
T	7	0	3	0	0	0	0	4	3	0	7	0	24
TOTAL	10	10	10	10	10	10	10	10	10	10	10	10	120

A ce stade, avant d'avancer dans les différents calculs, on peut dors et déjà écrire une séquence consensus commune aux 10 individus en s'appuyant sur le code IUPAC :

R	A + G	puRines
Y	T + C	pYrimidines
M	A + C	Groupe aMino
K	G + T	Groupe Keto (cétone)
W	A + T	Weak (faible)
S	G + C	Strong (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 Nucleotide

L'application de code IUPAC pour les 12 positions conduit à l'écriture suivante :

Position / Individu	1	2	3	4	5	6	7	8	9	10	11	12
1	T	T	C	C	C	G	A	A	G	G	C	A
2	C	T	T	C	G	G	G	T	T	G	T	A
3	C	T	T	C	G	G	A	T	T	G	T	A
4	T	T	T	C	G	G	A	T	C	G	T	A
5	T	T	C	C	C	G	A	A	G	G	G	A
6	T	C	C	C	C	G	A	A	G	G	G	A
7	T	C	C	C	C	G	A	A	G	G	T	A
8	T	T	C	C	G	G	A	A	G	G	T	A
9	T	T	C	C	G	G	A	T	G	G	T	A
10	C	T	C	C	G	G	G	A	T	G	T	A
Consensus	*Y*	*Y*	*Y*	*C*	*S*	*G*	*R*	*W*	*B*	*G*	*B*	*A*

Mais cette définition, vous l'avez remarqué, reste trop ambigüe et peu précise ; d'où l'utilité de poursuivre les calculs.

On peut utiliser une écriture plus précise en appliquant les conventions symboliques suivantes :

<G	Le nucléotide G est au début de la séquence du motif
A	Une seule lettre indique que le nucléotide est à la position donnée du motif
X	N'importe quel nucléotide est toléré à cette position
[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. Il est interdit à cette position. Jamais T à cette position, mais A ou C ou G sont possibles.
[CT] (2)	On peut avoir à cette position soit 2C soit 2T. Les deux parenthèses doivent toujours contenir des valeurs numériques (un entier). Le nucléotide est alors répété n fois consécutivement.
T (1,2)	Le nucléotide T est retrouvé (répété) entre une et deux fois.
-	Symbole ou élément qui sépare les résidus du motif
A>	Le nucléotide A est à la fin du motif

Exemple d'écriture :

< [AT]-G-X(3)-A-T > Le motif qui découle de cette écriture est

A-G-X-X-X-A-T OU **T**-G-X-X-X-A-T

Donc toutes les séquences qui possèdent un tel motif pourraient faire partie de la même "famille", donc de même fonction biologique.

Pour le cas de nos 10 individus, on peut écrire la séquence du motif comme suit :

Position / Individu	1	2	3	4	5	6	7	8	9	10	11	12
1	T	T	C	C	C	G	A	A	G	G	C	A
2	C	T	T	C	G	G	G	T	T	G	T	A
3	C	T	T	C	G	G	A	T	T	G	T	A
4	T	T	T	C	G	G	A	T	C	G	T	A
5	T	T	C	C	C	G	A	A	G	G	G	A
6	T	C	C	C	C	G	A	A	G	G	G	A
7	T	C	C	C	C	G	A	A	G	G	T	A
8	T	T	C	C	G	G	A	A	G	G	T	A
9	T	T	C	C	G	G	A	T	G	G	T	A
10	C	T	C	C	G	G	G	A	T	G	T	A

< [CT] - [CT] - [CT] - C - [CG] - G - [AG] -[AT] - [GT] - G - {A} - A >

Toujours est-il que cette écriture reste incomplète et manque toujours de précision, on calcule alors la fréquence rapportée au nombre total des séquences :

− Chaque valeur à l'intérieur du tableau est divisée par son propre total (24 pour A, 28 pour C, 44 pour G et 24 pour T) ; on obtient les fréquences *fi*

− Les totaux de chaque nucléotide seront divisés par le grand total, soit 120 : on obtient les fréquences *Fi*

La table des fréquences (*fi*) aura la forme suivante :

	Fréquences *fi*												Fréquences *Fi*
A	0/24	0/24	0/24	0/24	0/24	0/24	8/24	6/24	0/24	0/24	0/24	10/24	24/120 = 0,20
C	3/28	2/28	7/28	10/28	4/28	0/28	0/28	0/28	1/28	0/28	1/28	0/28	28/120 = 0,23
G	0/44	8/44	0/44	0/44	6/44	10/44	2/44	0/44	6/44	10/44	2/44	0/44	44/120 = 0,37
T	7/24	0/24	3/24	0/24	0/21	0/21	0/21	4/24	3/24	0/24	7/24	0/24	24/120 = 0,20

Ce qui donne, après calculs :

	fi												*Fi*
A	0	0	0	0	0	0	0,333	0,25	0	0	0	0,417	24/120 = 0,20
C	0,107	0,071	0,25	0,357	0,143	0	0	0	0,036	0	0,036	0	28/120 = 0,233
G	0	0,182	0	0	0,136	0,227	0,045	0	0,136	0,227	0,045	0	44/120 = 0,367
T	0,292	0	0,125	0	0	0	0	0,167	0,125	0	0,292	0	24/120 = 0,20

5. Table de poids de position

Cette étape sera la transformation de la table de fréquences en une table de poids. On obtient alors la table de pondération (Weight Matrix) ou table de poids position. Celle-ci est généralement construite en calculant le rapport des fréquences *fi/Fi*

	Rapport *fi/Fi*											
A : *fi/Fi = fi / 0,20*	0	0	0	0	0	0	1,665	1,25	0	0	0	2,085
C : *fi/Fi = fi / 0,233*	0,459	0,305	1,073	1,532	0,614	0	0	0	0,154	0	0,154	0
G : *fi/Fi = fi / 0,367*	0	0,496	0	0	0,371	0,618	0,123	0	0,371	0,618	0,123	0
T : *fi/Fi = fi / 0,20*	1,46	0	0,625	0	0	0	0	0,835	0,625	0	1,46	0

Une dernière transformation est effectuée en prenant le logarithme de la fréquence de chaque base à chaque position pour optimiser les différences contenues dans la table des fréquences.

ATTENTION : Le logarithme des valeurs nulles, qui est indéterminé, doit être remplacé, par convention, par une pénalité p égale à -10.

	Ln du Rapport *fi/Fi* : *Ln(fi/Fi)*											
A : Ln (*fi/Fi*)	-10	-10	-10	-10	-10	-10	0,510	0,223	-10	-10	-10	0,735
C : Ln (*fi/Fi*)	-0,779	-1,187	0,070	0,427	-0,488	-10	-10	-10	-1,871	-10	-1,871	-10
G : Ln (*fi/Fi*)	-10	-0,701	-10	-10	-0,991	-0,481	-2,096	-10	-0,991	-0,481	-2,096	-10
T : Ln (*fi/Fi*)	0,378	-10	-0,470	-10	-10	-10	-10	-0,180	-0,470	-10	0,378	-10

6. la séquence consensus

Il suffit de placer le nucléotide ayant la plus forte valeur de *Ln(fi/Fi)* à chaque position de l'alignement. On obtient alors une séquence consensus représentative du motif :

	Ln du Rapport *fi/Fi*											
A : Ln (*fi/Fi*)	-10	-10	-10	-10	-10	-10	**0,510**	**0,223**	-10	-10	-10	**0,735**
C : Ln (*fi/Fi*)	-0,779	-1,187	**0,070**	**0,427**	**-0,488**	-10	-10	-10	-1,871	-10	-1,871	-10
G : Ln (*fi/Fi*)	-10	**-0,701**	-10	-10	-0,991	**-0,481**	-2,096	-10	-0,991	**-0,481**	-2,096	-10
T : Ln (*fi/Fi*)	**0,378**	-10	-0,470	-10	-10	-10	-10	-0,180	**-0,470**	-10	**0,378**	-10
Consensus	**T**	**G**	**C**	**C**	**C**	**G**	**A**	**A**	**T**	**G**	**T**	**A**
Position	1	2	3	4	5	6	7	8	9	10	11	12

Remarquons que, comme déjà cité plus haut, les positions 4, 6, 10 et 12 ont bien conservé leurs nucléotides respectifs. La séquence qui résume au mieux les 10 séquences alignées est donc :

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

Le score de cette séquence sera calculé à partir de la matrice de poids position :