

Analyse de Séquences

M1 BIBS

Représentation et recherche de motifs

<https://github.com/dgautheret/M1BIBS-ana-seq>



V. 2018.1

Plan

- Représenter les motifs
- Estimer la performance d'une recherche de motifs
- Application à la détection de gènes
- Découverte de motifs inconnus
- Logiciels

motif = synthèse d'un ensemble de séquences homologues

- Identifier les résidus essentiels,
- Identifier les domaines fonctionnels
- Etablir des signatures pour rechercher de nouvelles instances du motif

Mais comment représenter ce qui est important?

eukaryotic TATA-box promoter sequences:

TCTATACAATGGC
ACTATATAATGGA
TGAATACATTGGG
TCTATACAATGCT
ACTATAATATTGC
TCTATATAATAGC

Consensus

- A partir d'un alignement, on détermine les résidus les plus fréquents à chaque position. Si la fréquence dépasse un certain seuil: séquence inclue dans le consensus. P. ex. consensus 75%:
- Représentation peu performante: faible spécificité / sensibilité

Position	1	2	3	4	5	6	7	8	9	10	11	12
% A	19.4	23.4	5.0	83.5	4.4	89.2	71.0	84.8	40.0	35.7	15.5	18.5
% C	22.7	34.0	11.0	1.3	3.3	0.8	0.8	2.9	3.4	14.0	36.5	37.0
% G	26.5	30.8	4.5	1.4	0.9	1.7	0.5	9.5	22.4	39.4	36.3	30.4
% T	31.4	11.7	79.5	13.9	91.4	8.4	27.7	2.8	34.2	10.8	11.7	14.1
Consensus			T	A	T	A	W	A	D	R		

Eukaryotic TATA-box promoter sequences (consensus à 75%)

Code IUPAC

W = A or T
S = C or G
R = A or G
Y = C or T
K = G or T
M = A or C
B = C, G, or T (not A)
D = A, G, or T (not C)
H = A, C, or T (not G)
V = A, C, or G (not T)
N = A, C, G, or T

Expressions régulières

- Un chaine de caractères décrivant un ensemble des séquences, avec des alternatives possibles à chaque position. c'est la méthode utilisée dans PROSITE. Exemple de descripteur PROSITE:
 - [AC] -x-V-x (4) -{ ED }
 - x: N'importe quel aa
 - []: choix entre plusieurs aa
 - { }: Tous, sauf les aa mentionnés
 - (x,y): Répétition x à y fois
 - ★ Semblables en principe, les langages utilisés dans Prosite et dans les expressions régulières Unix diffèrent dans les détails.
- ^ Le début d'une ligne
 - . Tout caractère (sauf newline)
 - \$ La fin d'une ligne
 - | Choix. A|B: A ou B
 - () groupement
 - [] Classe de caractères. [AGUC]: A,G,U ou C
 - \ Avant un caractère spécial
 - * 0 fois ou plus
 - + une fois ou plus
 - ? une fois ou zero
 - {n} exactement n fois
 - {n,} au moins n fois
 - {n,m} de n a m fois

Expressions régulières Unix:

Profil ou Matrice poids-position (PWM: Position Weight Matrix)

- Plus subtil que les consensus: Pour chaque position de l'alignement, on détermine la fréquence d'observation des différents résidus.
- Ceci est résumé dans un tableau qui donne pour chaque position les comptes ou fréquences des 20 a.a. (ou 4 bases)
- Une matrice de score est calculée à partir du tableau, selon la formule:
$$S_{b,i} = \log(F_{b,i} / F_b)$$
 - $F_{b,i}$: fréquence de b à la position i
 - F_b : fréquence dans b dans l'ensemble du génome analysé
- La recherche est effectuée en faisant glisser une fenêtre sur la séquence à analyser et en calculant le score total à chaque position de la fenêtre.

Exemple de PWM

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

A	60	40	10	20	20
C	5	10	29	22	15
G	2	20	31	10	16
T	5	2	2	20	21

Comptes (Nb,i)



Fréquences observées (Fb,i)



PWM = $\log(Fb,i / Eb,i)$

Eb,i = freq attendue (expected)
de la base b à la position i

A	0.5	0.3	-0.3	0.0	0.0
C	-0.6	-0.3	0.2	0.1	-0.1
G	-1.0	0.0	0.2	-0.3	-0.1
T	-0.6	-1.0	-1.0	0.0	0.1

Parcourir un génome avec une PWM

A	0.5	0.3	-0.3	0.0	0.0
C	-0.6	-0.3	0.2	0.1	-0.1
G	-1.0	0.0	0.2	-0.3	-0.1
T	-0.6	-1.0	-1.0	0.0	0.1



A A C C A C G G A A A C C A C G G A A A C C

+0,5 -0,3 +0,2 -0,3 +0,0

Score=-0.3

Limitations des PWM

- Pas de traitement des indels
- Mauvaise gestion des alignements pauvres en information
 - Problème de fréquences à zéro

Un PWM « creux »

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

A	6	4	0	2	2	0	1	1	2
C	0	0	3	1	0	6	0	3	0
G	0	2	3	1	0	0	2	2	0
T	0	0	0	2	4	0	3	0	4

$\log(0)$

$PWM = \log(Fb,i / Eb,i)$

0.6	0.4	#####	0.1	0.1	#####	-0	-0	0.1
#####	#####	0.3	-0	#####	0.6	#####	0.3	#####
#####	0.1	0.3	-0	#####	#####	0.1	0.1	#####

Manque d'information dans l'alignement

Un (trop petit) jeu de séquences d'entraînement :

TC t GGCTGGT	caaac-	GGA a	CCAA	gtccgtttcctgagaggt---	TTGG	TCC	CCTTCA	ACCAGCT a	CA
TG t GGCTGGT	caaac-	GGA a	CCAA	gtcaggtgttctgtgaggt--	TTGG	TCC	CCTTCA	ACCAGAC t	AT
TG t GGCTGGT	aaaac-	GGA a	CCAA	gtcaggtgttttgtgaggt--	TTGG	TCC	CCTTCA	ACCAGCT a	TG
TG c GGCTGGT	aaaaa-	GGA a	CCAC	atcaaccaggaaaaaggat---	TTGG	TCC	CCTTCA	ACCAGCC g	CA
TA t GGCTGGT	caaac-	GGA a	CCAA	gtccgtttccttagaggt---	TTGG	TCC	CCTTCA	ACCAGCT a	TT
AG t TGCTGGT	aaaac-	GGA a	CCAA	gtcgggtgttgcgagaggt--	TTGG	TCC	CTTTCA	ACCAGCT a	CT
TG t GGCTGGT	caaat-	GGA a	CCAA	gtcaggtgttctgcgaggt--	TTGG	TCC	CCTTCA	ACCAGCT a	CT

100% C

Autres scores = $\log(\text{obs}/\text{expected})$ = valeur arbitraire!

Toute autre séquence est elle vraiment impossible?

Que faire si on trouve un G ?

Solutions

1. Remplacer $\log(0)$ par une valeur arbitraire
(par ex -10)
 - Gros effet sur tous les résultats
2. Ajouter 1 à tous les comptages
 - Effet important sur petits alignements
3. Pseudocomptes
 - Principe: remplir les colonnes avec des comptages « raisonnables »
 - Exemple: la colonne c contient 7 C, on sait que T remplace souvent C. Insérons quelques T.
 - Il nous faut des matrices de substitution!

Pseudocomptes de Henikoff & Henikoff

$$N_{ai} = C * \sum_{b=A,T,G,C} P(b|i) * P(a|b)$$

Nb total de pseudocomptes (constante)

Nb de bases *a* ajoutées en colonne *i*

a remplacé par *b*

Avec exemple précédent:

Colonne *i* : 100% C

$P(C)=1$, autres=0

Nb de As dans col *i* = $C * 1 * P(A|C)$

Nb de Gs dans col *i* = $C * 1 * P(G|C)$ etc.

(on admet ici que $P=\text{fréquence}$)

PSI-Blast

(recherche itérative de profil à l'aide de PWM)

- **Principe**

- Une première séquence est recherchée dans une base de données
- Les séquences similaires significatives sont alignées sur la séquence requête.
- Un profil est construit
- Ce profil est recherché dans la banque de donnée pour collecter des séquences supplémentaires, etc.

- **Avantages et inconvénients**

- Excellent pour la recherche d'homologues éloignés.
- Si une séquence sans parenté avec la première est collectée accidentellement, celle-ci entraîne tous ses homologues avec elle au tour suivant. Le profil perd son sens.
- Les protéines multidomaines posent le même problème
- N'existe que pour les protéines

Output de Psi-Blast

Seuil de
E-value



<input type="checkbox"/> ubiquitin-conjugating enzyme 37 [Arabidopsis thaliana]	52.9	52.9	25%	9e-08	53%	NP_001325555.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/> PREDICTED: ubiquitin-conjugating enzyme E2 Q2 isoform X4 [Homo sapiens]	50.2	50.2	58%	9e-07	24%	XP_016878216.1	<input checked="" type="checkbox"/>	
<input type="checkbox"/> PREDICTED: ubiquitin-conjugating enzyme E2 Q2 isoform X3 [Homo sapiens]	50.2	50.2	58%	1e-06	24%	XP_016878215.1	<input checked="" type="checkbox"/>	
<input type="checkbox"/> PREDICTED: ubiquitin-conjugating enzyme E2 Q2 isoform X1 [Homo sapiens]	49.8	49.8	58%	1e-06	24%	XP_005254844.1	<input checked="" type="checkbox"/>	
<input type="checkbox"/> PREDICTED: ubiquitin-conjugating enzyme E2 Q2 isoform X5 [Homo sapiens]	46.0	46.0	55%	3e-05	23%	XP_016878217.1	<input checked="" type="checkbox"/>	
<input type="checkbox"/> PREDICTED: ubiquitin-conjugating enzyme E2 Q2 isoform X2 [Homo sapiens]	45.6	45.6	55%	4e-05	23%	XP_005254845.1	<input checked="" type="checkbox"/>	
<input type="checkbox"/> ubiquitin-conjugating enzyme E2 K isoform 5 precursor [Homo sapiens]	43.3	43.3	34%	6e-05	40%	NP_001299576.1	<input checked="" type="checkbox"/>	
<input type="checkbox"/> hypothetical protein, conserved [Leishmania donovani]	44.8	44.8	40%	8e-05	30%	XP_003861312.1	<input checked="" type="checkbox"/>	
<input type="checkbox"/> Uncharacterized protein CELE_C17D12.5 [Caenorhabditis elegans]	41.0	41.0	57%	8e-04	28%	NP_492956.2	<input checked="" type="checkbox"/>	

Run PSI-Blast iteration 3 with max 10000

Sequences with E-value WORSE than threshold									
Select: All None Selected:0		Yellow: sequences scoring below threshold on previous iteration							
		Description	Max score	Total score	Query cover	E value	Ident	Accession	
								Select for PSI blast	
								Used to build PSSM	
<input type="checkbox"/> PREDICTED: ubiquitin-conjugating enzyme E2 E3 isoform X4 [Homo sapiens]			36.7	36.7	16%	0.016	46%	XP_016858655.1	<input type="checkbox"/>
<input type="checkbox"/> PREDICTED: ubiquitin-conjugating enzyme E2 E3 isoform X3 [Homo sapiens]			36.7	36.7	16%	0.024	46%	XP_011508794.1	<input type="checkbox"/>
<input type="checkbox"/> ubiquitin-binding ESCRT-I subunit protein STP22 [Saccharomyces cerevisiae S288C]			37.1	37.1	39%	0.041	33%	NP_009919.3	<input type="checkbox"/>
<input type="checkbox"/> PREDICTED: receptor-like serine/threonine-protein kinase At1g78530 [Glycine max]			35.9	35.9	13%	0.045	60%	XP_006580801.1	<input type="checkbox"/>
<input type="checkbox"/> PREDICTED: ubiquitin-conjugating enzyme E2 E3 isoform X2 [Homo sapiens]			35.6	35.6	16%	0.051	46%	XP_016858655.1	<input type="checkbox"/>
<input type="checkbox"/> ubiquitin-conjugating enzyme E2 variant 1 isoform f [Homo sapiens]			35.2	35.2	85%	0.052	14%	NP_001244325.1	<input type="checkbox"/>
<input type="checkbox"/> ubiquitin-conjugating enzyme E2 variant 2 isoform 2 [Mus musculus]			35.2	35.2	85%	0.055	13%	NP_001152823.1	<input type="checkbox"/>
<input type="checkbox"/> ubiquitin-conjugating enzyme E2 variant 1 isoform 3 [Mus musculus]			34.8	34.8	85%	0.067	13%	NP_001298075.1	<input type="checkbox"/>

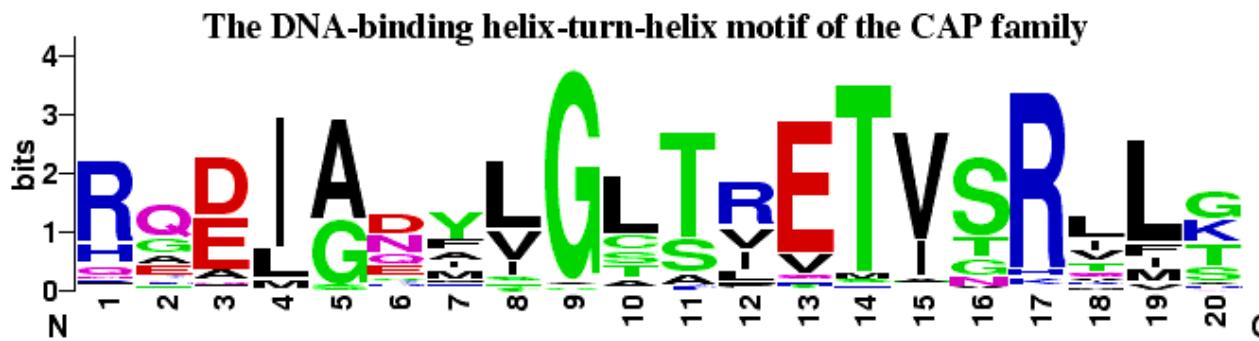
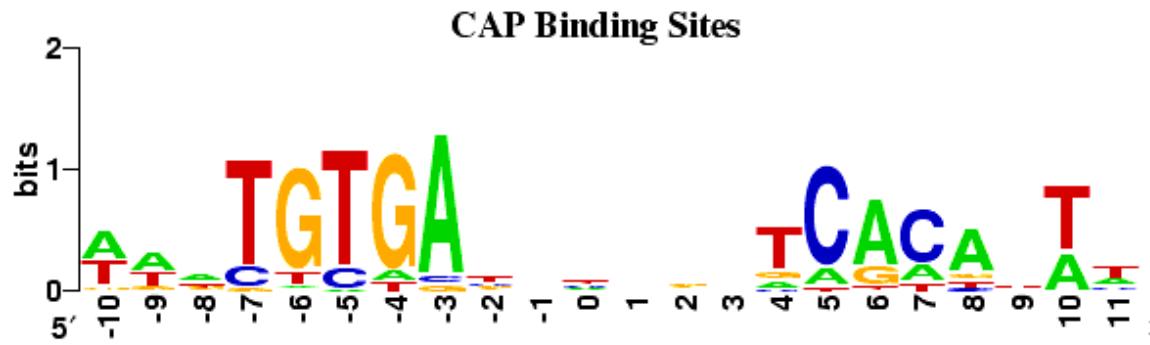
Sequence logos

(Schneider TD, Stephens RM. NAR. 1990)

Représentation en fréquence: sites de splicing



Représentation en logo:



Beaucoup mieux
qu'une fréquence!

Fait ressortir
régions
conservées/
variables

Entropie et contenu en information

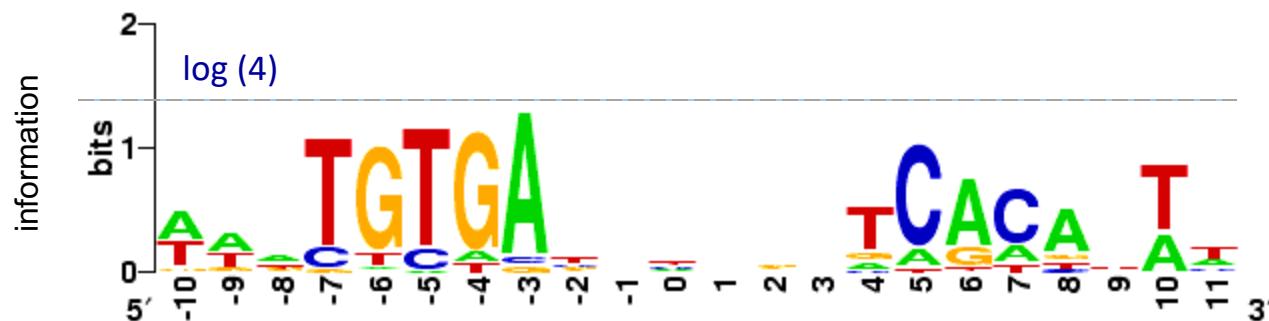
- Entropie de Shannon à la position i :

$$H_i = - \sum_{a=A,T,G,C} f_{a,i} \log(f_{a,i}).$$

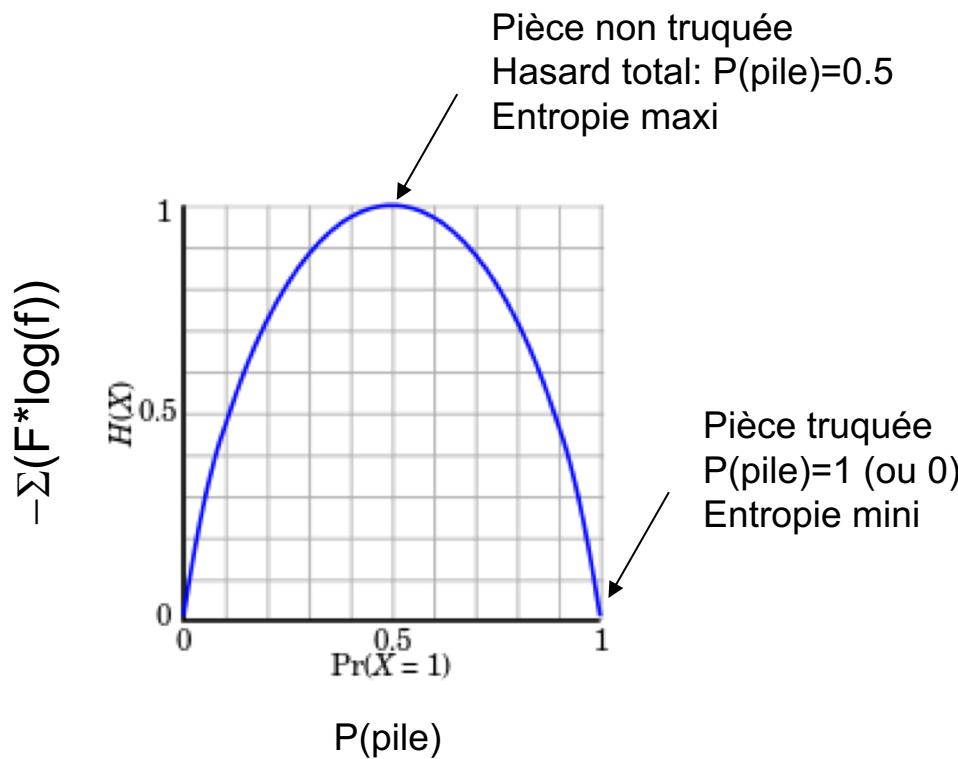
$f_{a,i}$: fréquence lettre a à la position i .

- Hauteur des lettres proportionnelle à:

$$f_{a,i} * (\log(4) - H_i) \quad = \text{information}$$



Pourquoi $-\sum(f \cdot \log(f))$?



...parce que cette fonction tend vers zero lorsque F tend vers 0 ou 1.

Exercice

- Récupérez le fichier *junctions.fa* contenant une série de jonctions 5' d'introns du chromosome 22 humain
- Combien y a-t-il de séquences?
- Utilisez le site Weblogo pour réaliser le logo de ce site.
- Comparez au motif « site de splicing » précédent

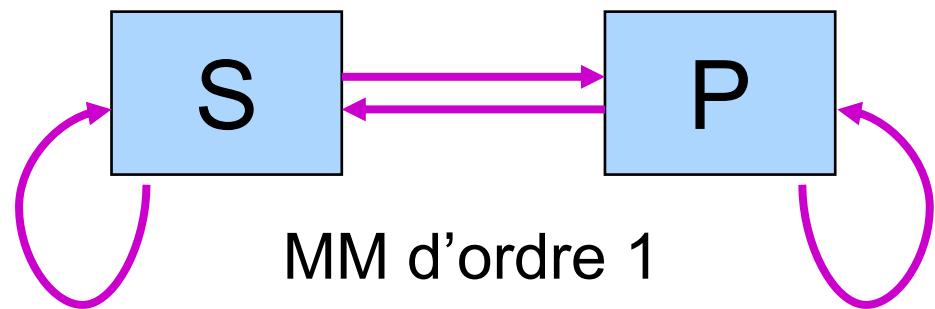
Motifs dans lesquels la position est indifférente = motif compositionnel

- Par exemple:
 - Reconnaître un gène d'une région intergénique
 - Reconnaître un exon d'un intron
 - Reconnaître un îlot CpG dans un génome
 - Reconnaître un domaine protéique intramembranaire d'un domaine soluble.

Dans tous ces cas: pas de motif de séquence avec bases ou aa attendus à des positions spécifiques

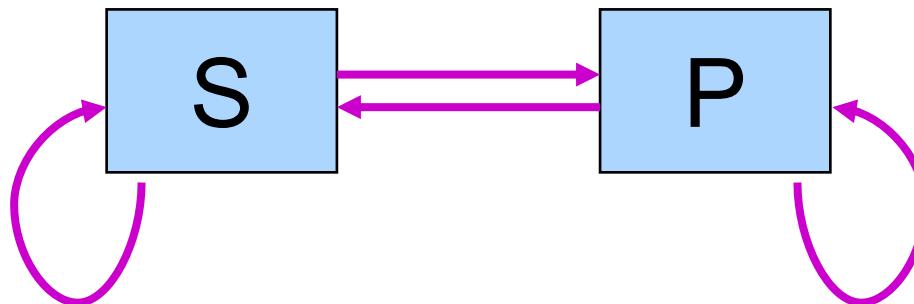
Chaînes de Markov (MM: Markov Models)

- Imaginons un climat à deux états:
 - P=Pluie
 - S=Soleil
- **S P S S P P P S P P P** -> temps demain ?
 - Quel est le temps le plus probable demain?
 - Solution: mesurer les probabilités de transition sur un ensemble d'entraînement (par ex tout une année), puis les appliquer à la dernière séquence observée
 - Ordre 1: P -> ?
 - Ordre 2: P S -> ?



Chaînes de Markov

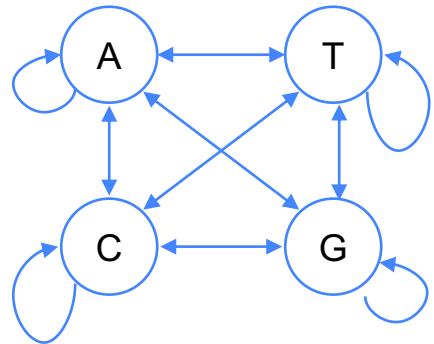
- *Chaine de Markov*: collection d'ETATS correspondant chacun à une observation, où le passage d'un état à l'autre (flèches) est associé à une probabilité.
- les probabilités de passage d'un état à l'autre sont appelées *probabilités de transition*.
- Le système a besoin d'une phase d'*entraînement* pour déterminer les probabilités de transition.



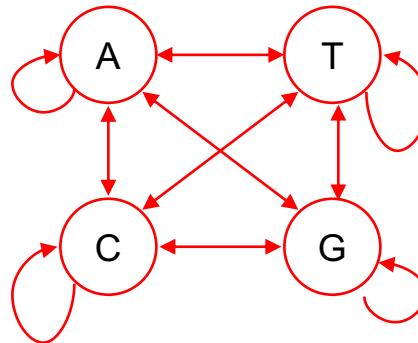
Modèles de Markov cachés (HMM)

- Cas où l'information que l'on cherche n'est pas un évènement de la chaîne.
- Par exemple découvrir le modèle “Eté” ou “Hiver”
- Il n'y a plus de correspondance directe entre les observations et les états. Par ex. l'observation « Pluie » peut se trouver dans un modèle ou dans l'autre. On dit alors que le modèle est **caché**

HMM et séquences biologiques



MM d'ordre 1
pour gène



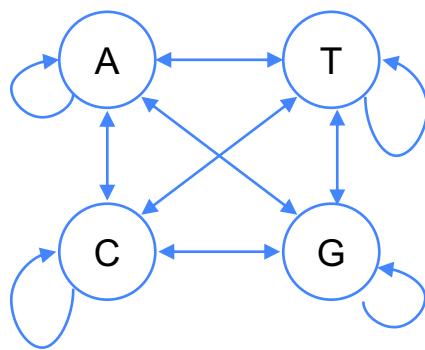
MM d'ordre 1
pour intergène

Séquence: CGACTTAGCACTTACGTAGCTGACGTACGCAT

Pour calculer $P(\text{séquence} == \text{gène})$

1. Extraire toutes les transitions de la séquence
2. Se reporter au MM gène pour obtenir les probas.
3. $P = \text{produit des probas}$

Mode « parcours »



intergène

gène

intergène

CATGCTAGCGATCTAGCTGACTAGGAGGGCGATTGCGACGGATTGATGAGTCGAT

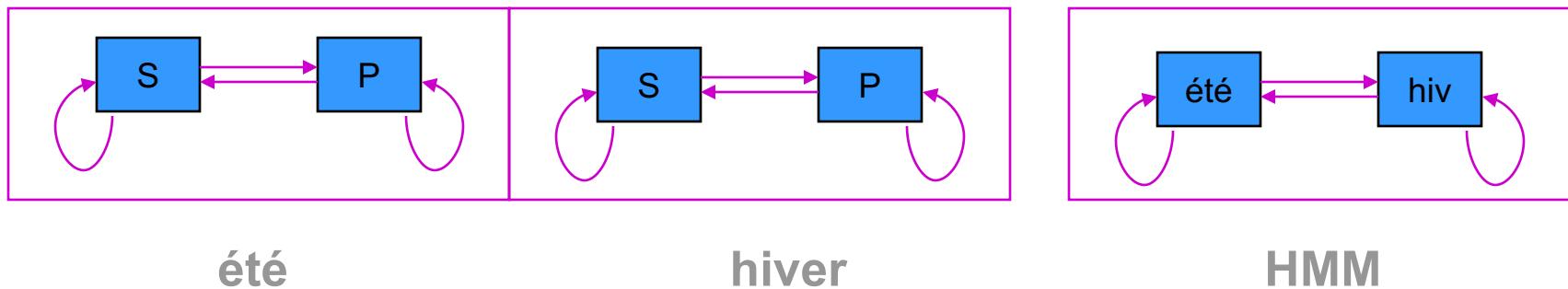
?

?

?

HMM: détecter le passage d'un modèle à l'autre

- Si l'on veut détecter le passage d'un modèle à l'autre, il faut ajouter à chaque état d'un modèle une probabilité de passer à un état de l'autre modèle.
- **S P S S P P P S P P P....** -> trouver les phases été et hiver dans la chaîne.
- Dans ce cas, il faut entraîner deux MM (été et hiver) et évaluer en plus les transitions été/hiver:



Recherche avec HMM

- L'algorithme de Viterbi est utilisé pour découvrir la suite d'état cachés la plus probable dans une séquence donnée.
 - Algorithme de programmation dynamique

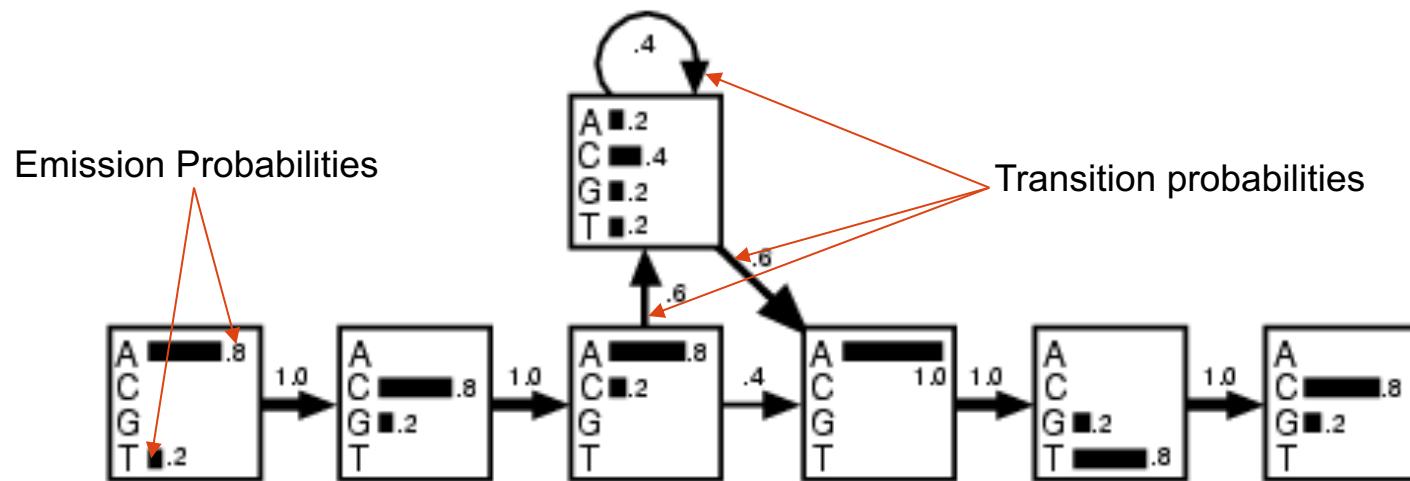
HMM et séquences biologiques

- Pour les motifs où la **succession** des nt ou aa est importantes
 - CpG (2nt) (promoteurs)
 - Codons (3nt) (gènes)
 - Régions hydrophobes (segments intermembranaires)
- **Base d'entraînement**= ensemble de séquences de la même famille:
 - Exons, introns, promoteurs, îlots CpG

Profile-HMM = un HMM position-spécifique

- Un HMM de taille définie où chaque position a ses propres probas de transition et d'identité
- Typiquement produit à partir d'un alignement multiple
- Combine les avantages des PWM et des HMM: les positions importantes sont identifiées, les insertions/déletions sont autorisées

Exemple de profil-HMM



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

HMMER

S. Eddy ; hmmer.janelia.org

Deux commandes: hmmbuild, hmmsearch:

First build a profile HMM from the alignment. The following command builds a profile HMM from the alignment of 50 globin sequences in **globins50.msf**:

> **hmmbuild globin.hmm globins50.msf**

Then use the globin hmm to search for globin domains in the *Artemia* globin sequence in **Artemia.fa**:

> **hmmsearch globin.hmm Artemia.fa**

HMMER output

Parsed for domains:

Sequence	Domain	seq-f	seq-t	hmm-f	hmm-t	score	E-value
S13421	7/9	932	1075	..	1 143 []	76.9	7.3e-24
S13421	2/9	153	293	..	1 143 []	63.7	6.8e-20
S13421	3/9	307	450	..	1 143 []	59.8	9.8e-19
S13421	8/9	1089	1234	..	1 143 []	57.6	4.5e-18
S13421	9/9	1248	1390	..	1 143 []	52.3	1.8e-16
S13421	1/9	1	143	[.	1 143 []	51.2	4e-16
S13421	4/9	464	607	..	1 143 []	46.7	8.6e-15
S13421	6/9	775	918	..	1 143 []	42.2	2e-13
S13421	5/9	623	762	..	1 143 []	23.9	6.6e-08

Alignments of top-scoring domains:

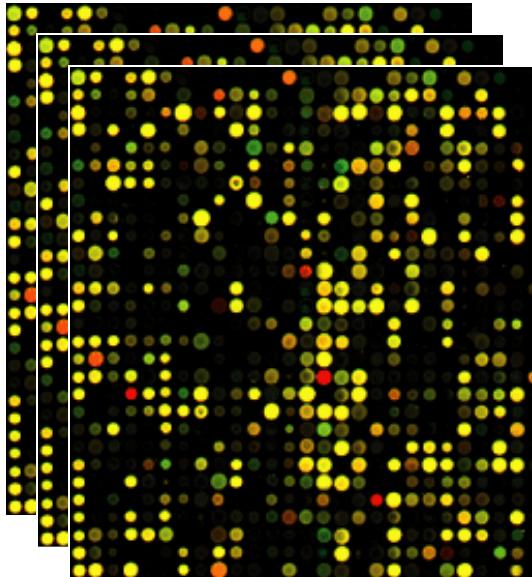
S13421: domain 7 of 9, from 932 to 1075: score 76.9, E = 7.3e-24
*->ekalvksvwgkveknveevGaeaLerllvvvPetkryFpkFkdLss
+e a vk+ w+ v+ ++ vG +++ l++ +P+ +++FpkF d+
S13421 932 REVAVVKQTWNLVKPDLMGVGMRIFKSLFEAFPAYQAVFPKFSDVPL 978

adavkgsakvkahgkkVltaulgavkkld...lkgalakLselHaqklr
d++++++ v +h V t+l++ ++ ld++ +l+ ++L+e H+ lr
S13421 979 -DKLEDTPAVGKHSISVTTKLDELIQTLDEpanALLARQLGEDHIV-LR 1026

vdpenfkllsevllvvlaeklgkeftpevqaalekllaavataLaakYk<
v+ fk +++vl+ l++ lg+ f+ ++ +++k++++++ +++ +
S13421 1027 VNKPMPFKSFGKVLVRLLENDLGQRFSSFASRSWHKAYDVIVEYIEGLQ 1075

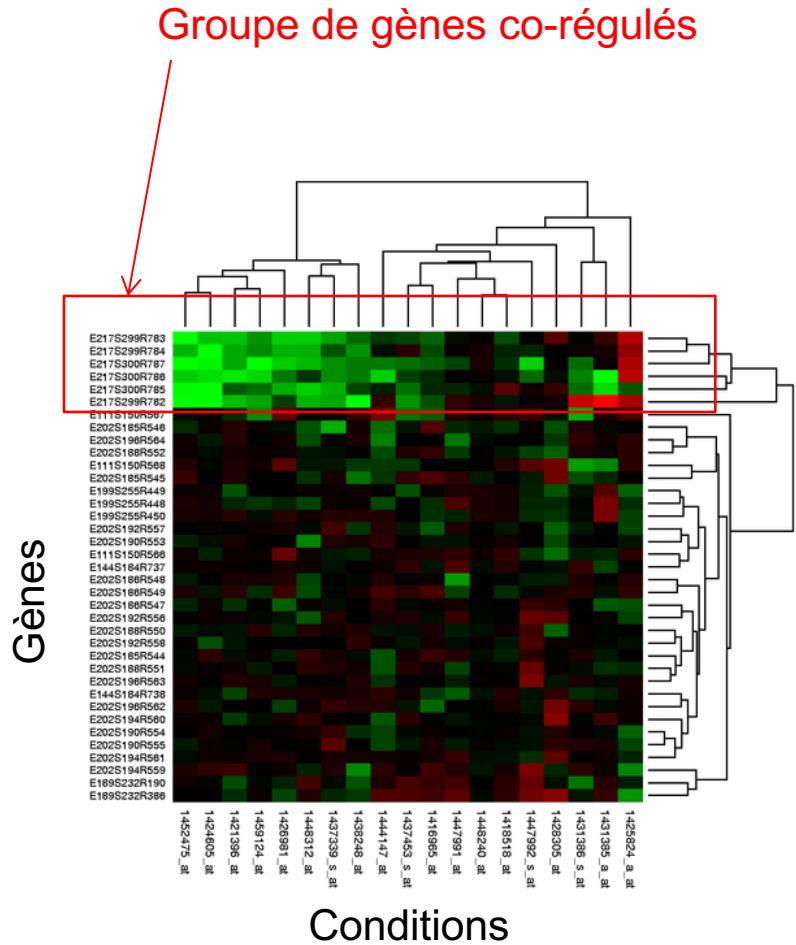
Découvrir des motifs inconnus

Exemple: groupes de gènes co-régulés identifiés par analyse de transcriptome

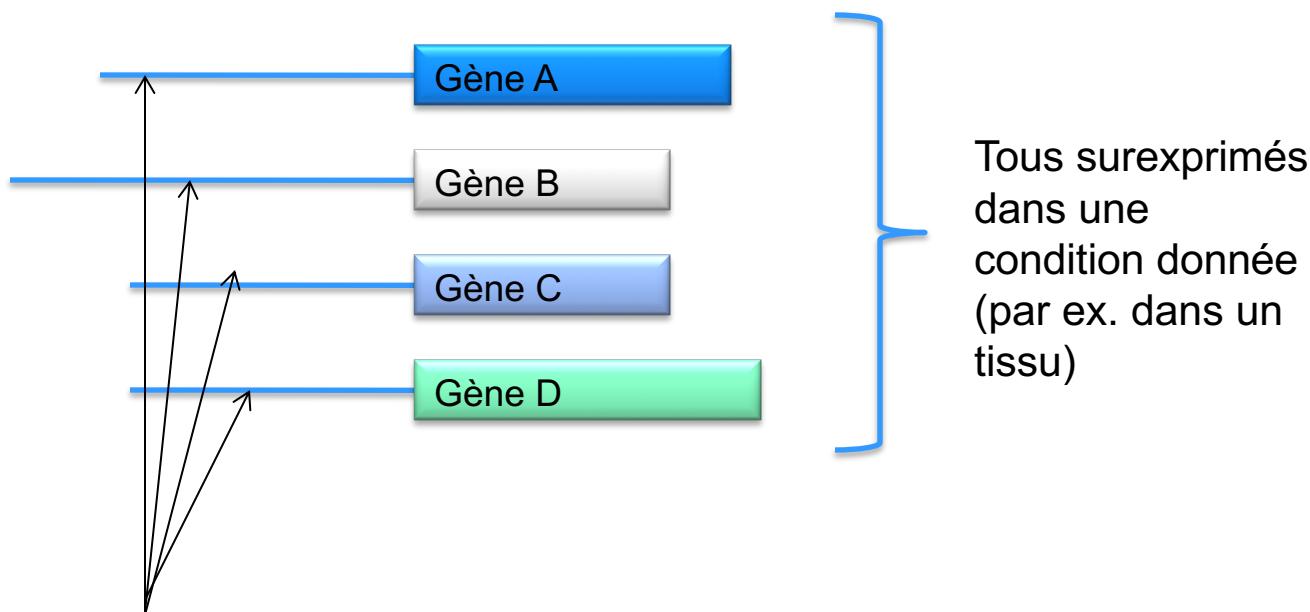


Puces à ADN ->
expression des gènes
dans différentes conditions

Clustering



Découverte de motifs régulateurs

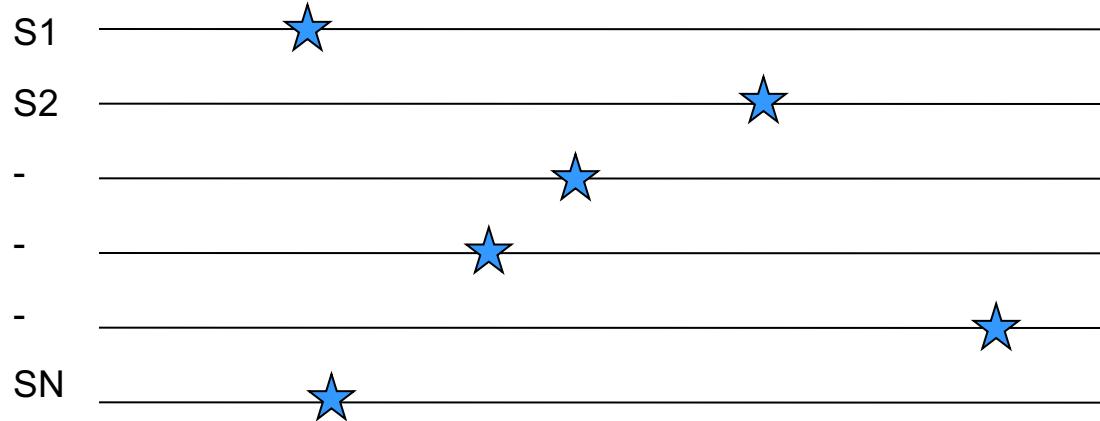


Séquence régulatrice commune?

(= Site de fixation pour un Facteur spécifique du tissu)

Recherche de motifs enrichis dans un ensemble de séquences

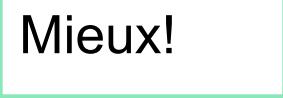
- Principe:
 - N séquences



But: Découvrir ☆

Recherche par comptage de mots

- Idée: rechercher des mots de k lettres surreprésentés dans la région d'intérêt
- Mais surreprésentés par rapport à quoi?
 - Une composition uniforme 25%A/25%T/25%G/25%C?
 - La composition du génome étudié?
 - La composition des régions étudiées?



Mieux!

Combien de fois s'attend-on à trouver un mot w ?

- $Ew = Pw \times T$

Ew : nombre d'occurrences attendues

Pw : probabilité d'observer w à chaque position

T : nombre de positions

(= longeur de la chaîne – taille du mot)

Pw

- Produit des fréquences des bases
- Corrigé par l'autocorrélation:

CCTAA

Sans autocorrélation: $P_w =$
produit des fréquences des bases

CCT CCTCC

Autocorrélation: P_w
nécessite correction

Comment donner un score au nombre d'occurrences observées?

- Ratio
- Z-score / distribution normale
- Log de vraisemblance
- Distribution binomiale

Ratio (ou log-ratio)

- $r = C_w / E_w$

C_w : nombre d'occurrence observé de w

E_w : nombre d'occurrences attendu de w

- Facile, mais surestime l'importance des mots rarement attendus

Log de vraisemblance

- $K = F_w * \log (F_w / P_w)$

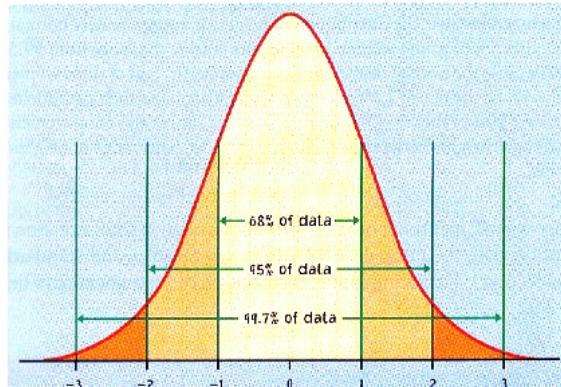
F_w : fréquence observée de w

P_w : fréquence attendue de w

- Valeur positive/négative pour motif enrichi/appauvri
- Ne donne pas une probabilité

Z-score

- $Z = (C_w - E_w) / S_w$



From: <http://www.sci.sdsu.edu/class/psychology/psy271/Weeks/psy271week06.htm>

C_w : nombre d'occurrences observé de w

E_w : nombre d'occurrences attendu de w

S_w : variance

- Peut se traduire en P-value grâce à la distribution normale, mais:
- Nécessite un grand jeu de données
- La distribution normale/gaussienne est approximative, surtout dans les valeurs extrêmes

Loi binomiale et P-value

- Calcul exact de la probabilité d'observer C_w fois ou plus le mot w
- Faiblesse: ne tient pas compte des mots chevauchants (autocorrélés)

Exactement C_w fois

$$P(X = C_w) = \frac{T!}{C_w!(T - C_w)!} p^{C_w} (1 - p)^{T - C_w} = C_T^{C_w} p^{C_w} (1 - p)^{T - C_w}$$

Plus de C_w fois

$$P(X \geq C_w) = \sum_{i=C_w}^T \frac{T!}{i!(T - i)!} p^i (1 - p)^{T - i}$$

Moins de C_w fois

$$P(X \leq C_w) = \sum_{i=0}^{C_w} \frac{T!}{i!(T - i)!} p^i (1 - p)^{T - i}$$

Programme utilisant les comptages de mots: RSA-tools

J. Van Helden: rsat.bigre.ulb.ac.be/rsat/

Extrait de la doc:

5.2.1 Counting word occurrences and frequencies

Try the following command:

```
oligo-analysis -v 1 -i Escherichia_coli_K12_start_codons.wc \
-format wc -l 3 -1str
```

Output:

```
;seq identifier exp_freq occ exp_occ occ_P occ_E occ_sig rank ovl_occ forbocc
acgtgc acgtgc|gcacgt 0.0002182431087 16 2.46 8.4e-09 1.7e-05 4.76 1 2 76
cccacg cccacg|cgtaaa 0.0001528559297 11 1.72 2e-06 4.2e-03 2.37 2 0 55
acgtgg acgtgg|ccacgt 0.0002257465554 13 2.54 2.8e-06 5.9e-03 2.23 3 1 65
cacgtg cacgtg|cacgtg 0.0001299168211 10 1.46 3.3e-06 6.8e-03 2.17 4 0 100
cgcacg cgcacg|cgtaaa 0.0001322750472 10 1.49 3.8e-06 8.0e-03 2.10 5 0 50
cgtata cgtata|tatacg 0.0005113063008 17 5.76 0.00011 2.2e-01 0.65 6 1 85
agagat agagat|atctct 0.0006913890231 19 7.78 0.00047 9.8e-01 0.01 7 0 95
```

Découvrir un motif enrichi par Expectation Maximization (EM)

P (Position Frequency Matrix) de taille W pour stocker le motif en cours de recherche

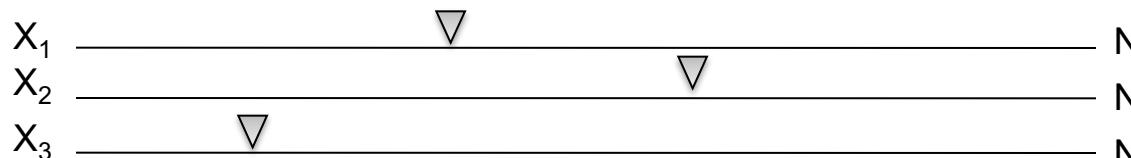
	1	2		W
A				
C				
G				
T				

Permet de calculer la probabilité d'une séquence, avec motif à une position déterminée

Matrice P_0 pour stocker les fréquences de fond (=attendues) des 4 bases

p_A
p_C
p_G
p_T

Séquences à analyser de taille N



∇ Positions du motif à découvrir

Calcul Matrice Z

$Z_{i,j}$ = Proba que le motif commence en position i dans la séquence j

	1	2	3	4	...	N-W
seq1						
seq2						
seq3						

Example: Estimating Z

$$X_i = \text{G C T G T A G}$$

$$p = \begin{matrix} & 0 & 1 & 2 & 3 \\ \text{A} & 0.25 & 0.1 & 0.5 & 0.2 \\ \text{C} & 0.25 & 0.4 & 0.2 & 0.1 \\ \text{G} & 0.25 & 0.3 & 0.1 & 0.6 \\ \text{T} & 0.25 & 0.2 & 0.2 & 0.1 \end{matrix}$$

$$Z_{i1} = [0.3 \times 0.2 \times 0.1 \times 0.25 \times 0.25 \times 0.25 \times 0.25]$$

$$Z_{i2} = 0.25 \times [0.4 \times 0.2 \times 0.6 \times 0.25 \times 0.25 \times 0.25]$$

⋮

Exemple avec un motif de taille W=3 et des séquences de taille N=7

From Mark Craven « Learning sequence motifs using expectation Maximization (EM) » 2002

EM: initialisations

1

Initialiser P en prenant toutes les positions de toutes les séquences et P_0 en prenant les fréquences de chaque base



2

A l'aide de P et P_0 , calculer la probabilité de chaque séquence pour chaque position de P. Certaines positions donnent une probabilité plus élevée = matrice Z

Matrice P

	1	2	W
A			
C			
G			
T			



X_1

↓ ↓ ↓ ↓ ↓ ↓ Une proba pour chaque position

X_2

Une proba pour chaque position

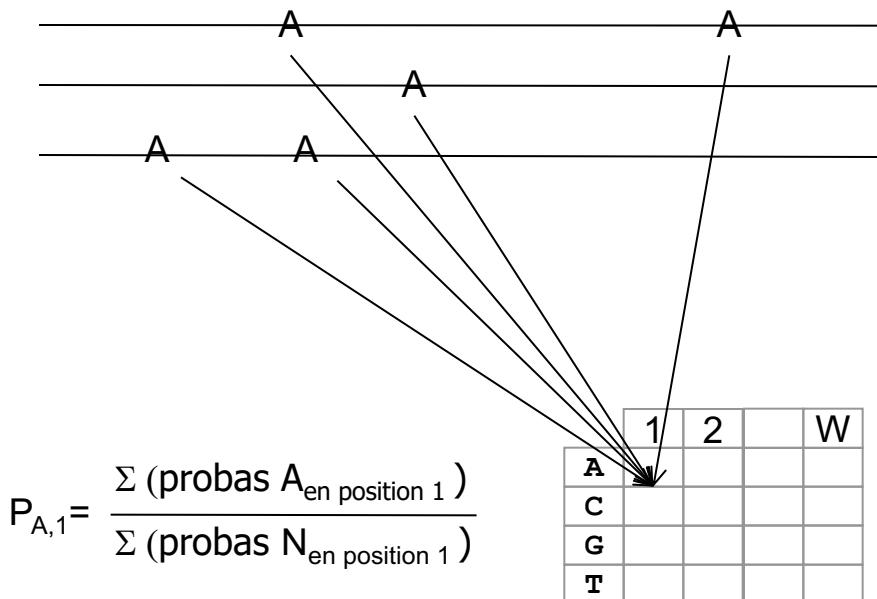
X_3

Une proba pour chaque position

Recalcul de P

3

Recalculer P en prenant dans Z les probabilités d'avoir chaque caractère c en position i . Recalculer P_0 en prenant 1-probabilité.



Example: Estimating p

$$\begin{matrix} \textbf{A} & \textbf{C} & \textbf{A} & \textbf{G} & \textbf{C} & \textbf{A} \\ \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ Z_{1,1} = 0.1, Z_{1,2} = 0.7, Z_{1,3} = 0.1, Z_{1,4} = 0.1 \end{matrix}$$

$$\begin{matrix} \textbf{A} & \textbf{G} & \textbf{G} & \textbf{C} & \textbf{A} & \textbf{G} \\ \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ Z_{2,1} = 0.4, Z_{2,2} = 0.1, Z_{2,3} = 0.1, Z_{2,4} = 0.4 \end{matrix}$$

$$\begin{matrix} \textbf{T} & \textbf{C} & \textbf{A} & \textbf{G} & \textbf{T} & \textbf{C} \\ \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ Z_{3,1} = 0.2, Z_{3,2} = 0.6, Z_{3,3} = 0.1, Z_{3,4} = 0.1 \end{matrix}$$

$$p_{A,1} = \frac{Z_{1,1} + Z_{1,3} + Z_{2,1} + Z_{3,3} + 1}{Z_{1,1} + Z_{1,2} \dots + Z_{3,3} + Z_{3,4} + 4}$$

Pseudocomptes

Itération...

- 1 Initialiser P en prenant toutes les positions de toutes les séquences et P_0 en prenant les fréquences de chaque base
- 2 A l'aide de P et P_0 , calculer la probabilité de chaque séquence pour chaque position de P . Certaines positions donnent une probabilité plus élevée = matrice Z
- 3 Recalculer P en prenant dans Z les probabilités d'avoir le caractère c en position i . Recalculer P_0 en prenant 1-probabilité.

Estimation Step

Réitérer jusqu'à ce que variation de $p < \varepsilon$

EM converge vers un minimum local – sensible aux valeurs initiales de p

The MEME web server

MEME Suite Menu
Submit A Job
Documentation
Downloads
User Support
Alternate Servers
Authors
Citing

 **MEME**
Multiple Em for Motif Elicitation

Version 4.9.0

Data Submission Form

Required

Your e-mail address:

Re-enter e-mail address:

Please enter the **sequences** which you believe share one or more motifs. The sequences may contain no more than **60000 characters** total in any of a large number of **formats**.

Enter the **name of a file** containing the sequences here:
 Aucun fichier sélectionné.

or

the **actual sequences** here (**Sample Protein Input Sequences**):

How do you think the occurrences of a single motif are **distributed** among the sequences?
 One per sequence
 Zero or one per sequence
 Any number of repetitions

MEME will find the optimum **width** of each motif within the limits you specify here:

6	Minimum width (>= 2)
50	Maximum width (<= 300)

Maximum **number of motifs** to find

Utilisation de MEME

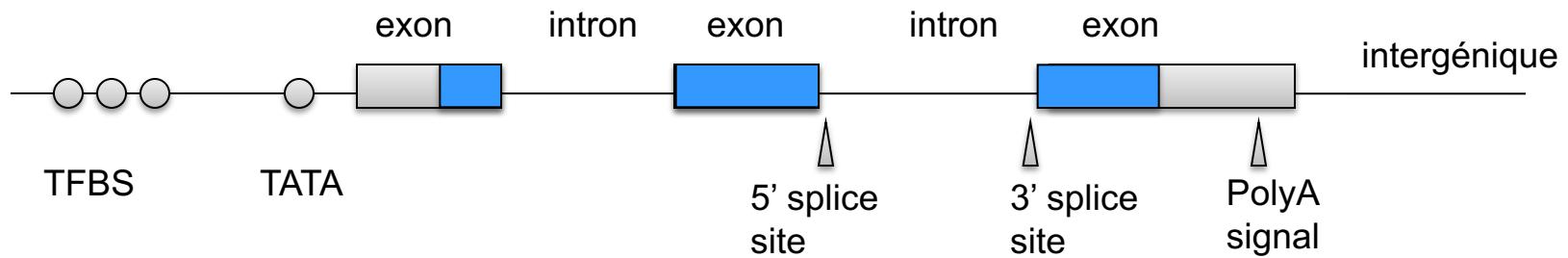
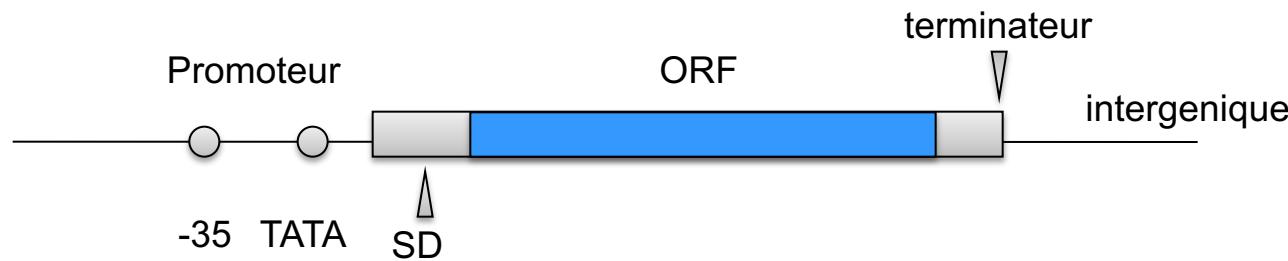
- MEME converge toujours vers la même solution
- Mais cette solution est très sensible aux paramètres de départ
 - Taille du motif
 - Nombre de motifs autorisés par séquence (0,1,N)
- Ne pas hésiter à essayer de nombreuses combinaisons de paramètres.

TD: recherche de motif dans la partie 3' des mRNA

- Récupérer fichier fasta contenant 250 extrémités de cDNA humains
 - [human-cDNA-tips.fa](#)
 - (40nt, cDNA pris au hasard)
- Analyse avec MEME
- Analyse avec RSATools
- Conclusion

Application à la detection de gènes

Déterminants des gènes microbiens et eucaryotes



La recherche de motifs seuls est insuffisante

TATA box, Séquence SD, sites d'épissage, TFBS...

Nombreux faux positifs: spécificité faible

- Même si les motifs étaient parfaitement conservés, ils sont peu spécifiques.
- P. ex: le site d'epissage consensus AxGT(A/G)xG est observé 1000 fois par segment de 100kb .
- Même une combinaison de motifs (TATA box + jonctions intron-exon + signal de polyadenylation) donnerait trop de FP

Open Reading Frames / cadres de lecture ouverts

- On trouve en moyenne un ORF de 150 nt (la taille typique d'un ORF) tous les kilobases, alors qu'il n'en existe en fait qu'un tous les 10kb dans les génomes de vertébrés.
- 9 Faux Positifs pour un Vrai Positif !

Les motifs de composition

- Motifs ne s'exprimant pas par une séquence consensus spécifique, mais par un biais de séquence.
 - Biais de GC
 - Biais de codon.

Les îlots CpG (vertébrés)

- Ilôts CpG : zones riches en dinucléotide CG
- Régions 5' des gènes (promoteur et exons 1-2)
- Fréquence attendue du dinucléotide CpG chez l'homme
= 4% (0.21×0.21), mais fréquence observée: un cinquième de cette valeur. Pourquoi?
 - Méthylation naturelle des CpG et réparation en TpG par déamination
 - Au niveau du promoteur: protection des CpG. Donc Fréquence normale.
- Typiquement 1-2kb de longueur. Environ 70% G+C (contre 40% dans le reste du génome humain)
- Les îlots CpG sont associés à tous les gènes housekeeping (constitutifs) et à 40% des autres gènes

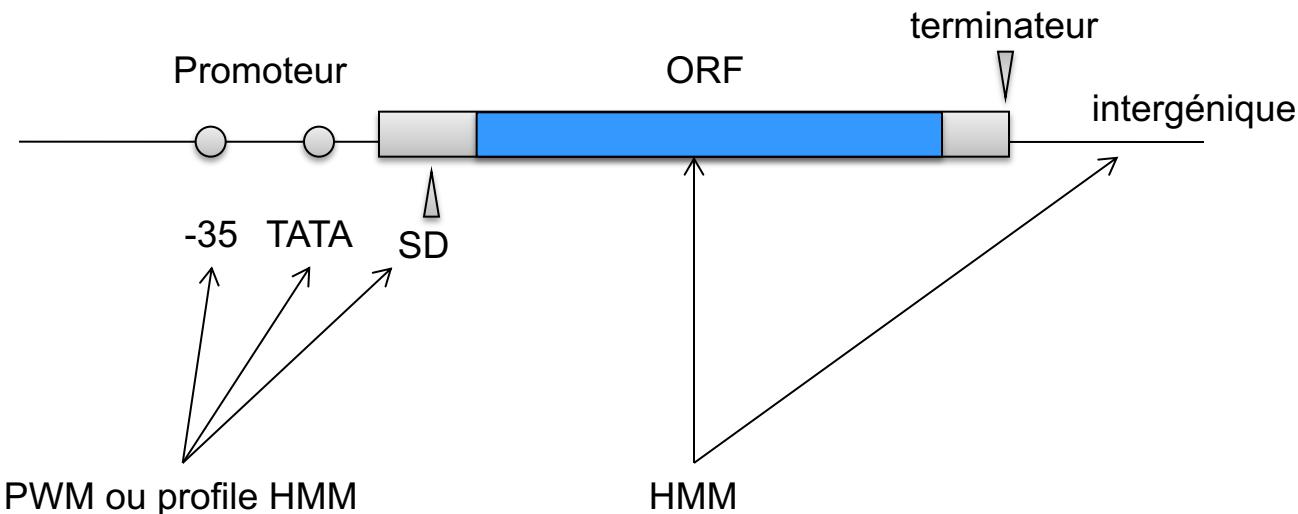
Biais d'usage de codons

- La composition en codons des ORF ne dépend pas seulement de la composition du génome: biais de codon
- Raisons:
 - Composition des protéines en aa
 - ARNt disponibles
 - Préférence pour les codons purine-N-pyrimidines
 - Préférence pour les codon/anticodons riches en GC qui minimisent les erreurs de traduction
- Ce biais est propre à l'espèce
- Il existe aussi des biais propres à certains gènes.
 - Généralement les gènes les plus exprimés sont les plus biaisés, les codons les plus utilisés étant ceux pour lesquels les ARNt sont les plus nombreux (codons sélectionnés pour une plus grande vitesse de traduction)

Capturer les fréquences de codons

- Modèle HMM à 2 ou 3 états
 - Procaryotes: codant / intergénique
 - Eucaryotes: Exons codants/introns et UTR/intergénique
 - Parfois état supplémentaire pour régions riches en CpG dans la partie 5' des gènes
- Beaucoup plus efficace que la recherche de motif

Modélisation des gènes microbiens



Annotation de gènes microbiens: Genemark

HMMs pour codant, non-codant et RBS.

Results of GeneMark.hmm predictions for 10 complete bacterial genomes*

Genome	Genes annotated	Genes predicted	Annotated genes predicted by		Correct 5' end prediction of annotated genes (%)	Potential new genes (%)
			GeneMark.hmm (%)	GeneMark (%)		
A.fulgidus	2407	2530	98.0	97.3	73.1	15.1
B.subtilis	4101	4384	97.2	97.3	77.5	9.8
E.coli	4288	4440	97.3	97.3	75.4	8.2
H.influenzae	1718	1840	96.2	96.2	86.7	10.2
H.pylori	1566	1612	95.6	95.6	79.7	8.7
M.genitalium	467	509	98.3	98.3	78.4	17.3
M.jannaschii	1680	1841	99.2	99.2	72.7	12.9
M.pneumoniae	678	734	95.9	95.9	70.1	13.6
M.thermoauto	1869	1944	97.5	97.5	70.9	8.6
Synechocystis	3169	3360	98.5	98.5	89.6	9.4
Average	21943	23194	97.3	97.3	78.1	10.4

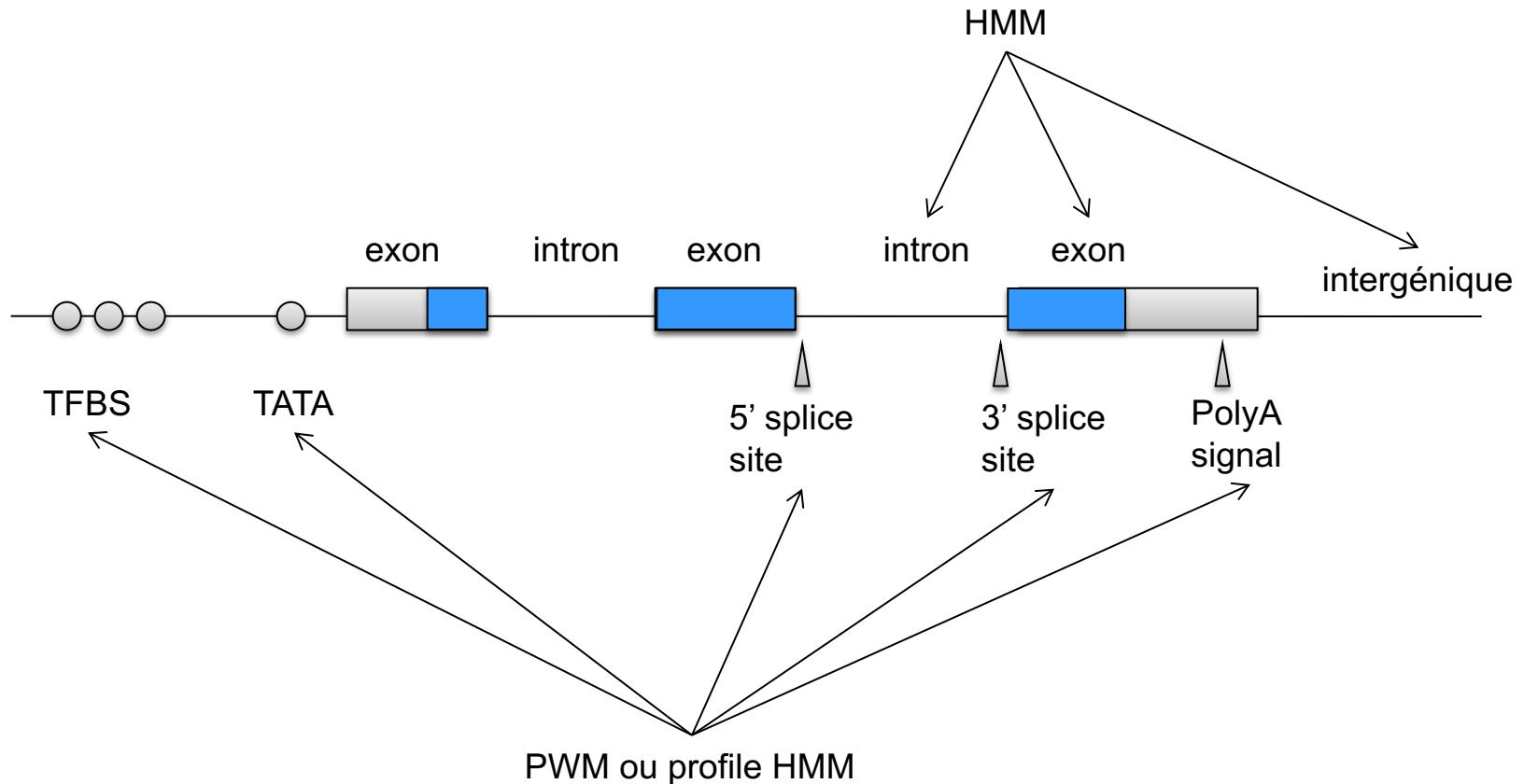
The second and third columns show the number of genes annotated in GenBank and the number of genes predicted, respectively.

The "Annotated genes predicted" column presents the percentage of annotated genes which were predicted by GeneMark and GeneMark.hmm. The "Correct 5' end prediction of annotated genes" column shows the percentage of genes whose starts were predicted exactly.

"Potential new genes" is the fraction of predicted genes for which no annotated analog was found. All measures are expressed in percent.

* Reference: A. Lukashin and M. Borodovsky, GeneMark.hmm: new solutions for gene finding, NAR, 1998, Vol. 26, No.4, pp 1107-1115.

Modélisation des gènes eucaryotes



Gènes eucaryotes

- GENSCAN, Fgenesh, Eugene
- Utilisent HMM. Plusieurs modèles sont employés pour exons, introns, promoteurs, etc. Sensibilité et Spécificité autour de 80% pour les exons correctement prédits. Beaucoup moins bon pour la prédiction de gènes complets.

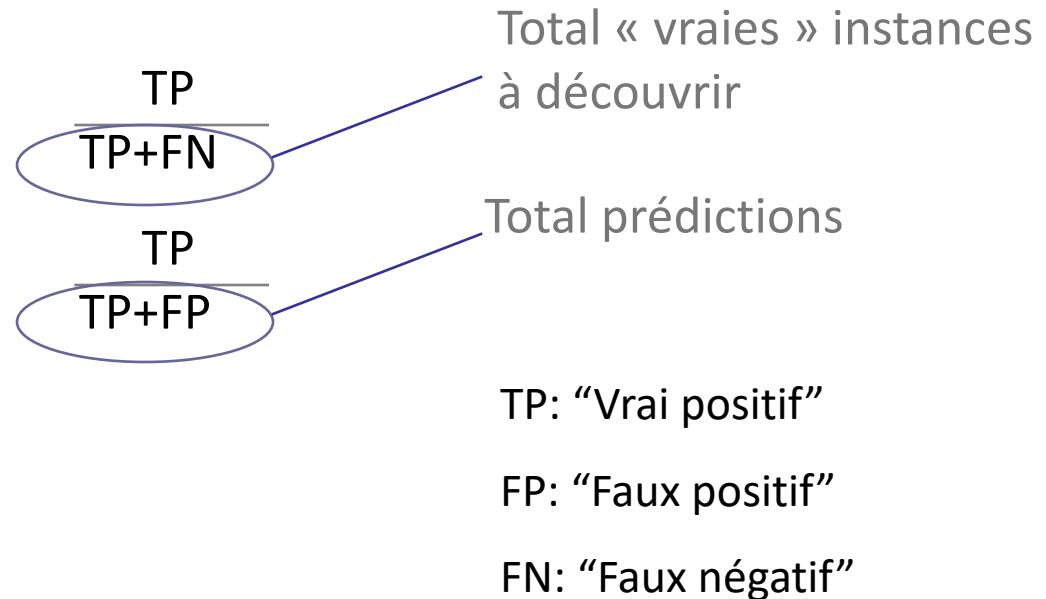
Estimer la performance d'un prédicteur

Sensibilité et Spécificité

- *Sensibilité*: La capacité à détecter les vraies instances de l'objet recherché (« vrais positifs »).
- *Spécificité*: La capacité à rejeter les fausses instances (« faux positifs »).

Sensibilité: $SN =$

Spécificité : $SP =$



La courbe ROC*

*Receiver Operator Characteristic

