

Learning Sequence Motif Models Using Expectation Maximization (EM)

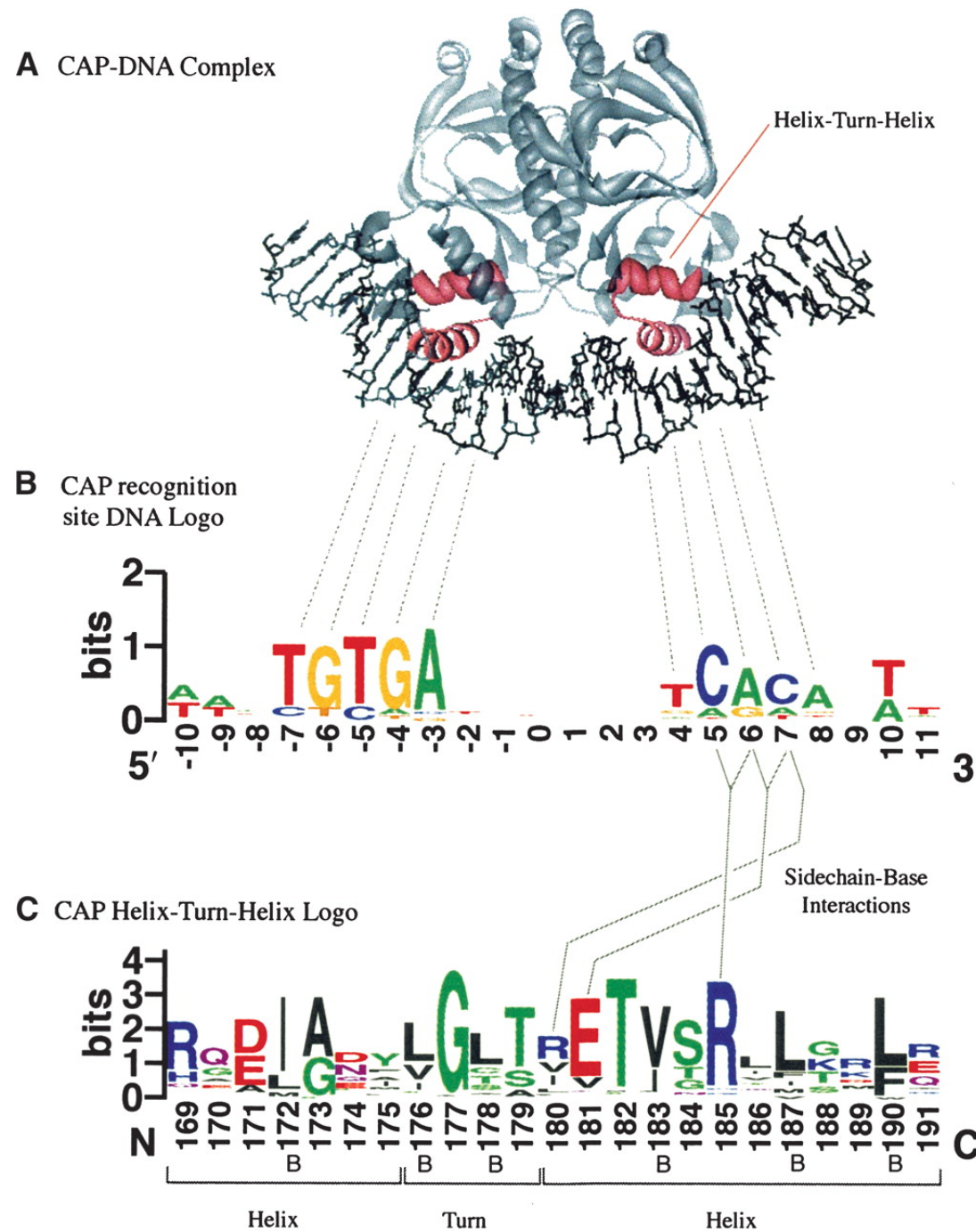
Juliana Silva Bernardes

Sequence Motifs

what is a sequence *motif*?

- A subsequence (substring) that occurs in multiple sequences with a biological importance.
- Motifs can be totally constant or have variable elements.
- Protein Motifs often result from structural features.
- DNA Motifs (regulatory elements)
 - Binding sites for proteins
 - Short sequences (5-25)
 - Inexactly repeating patterns

Sequence Motifs

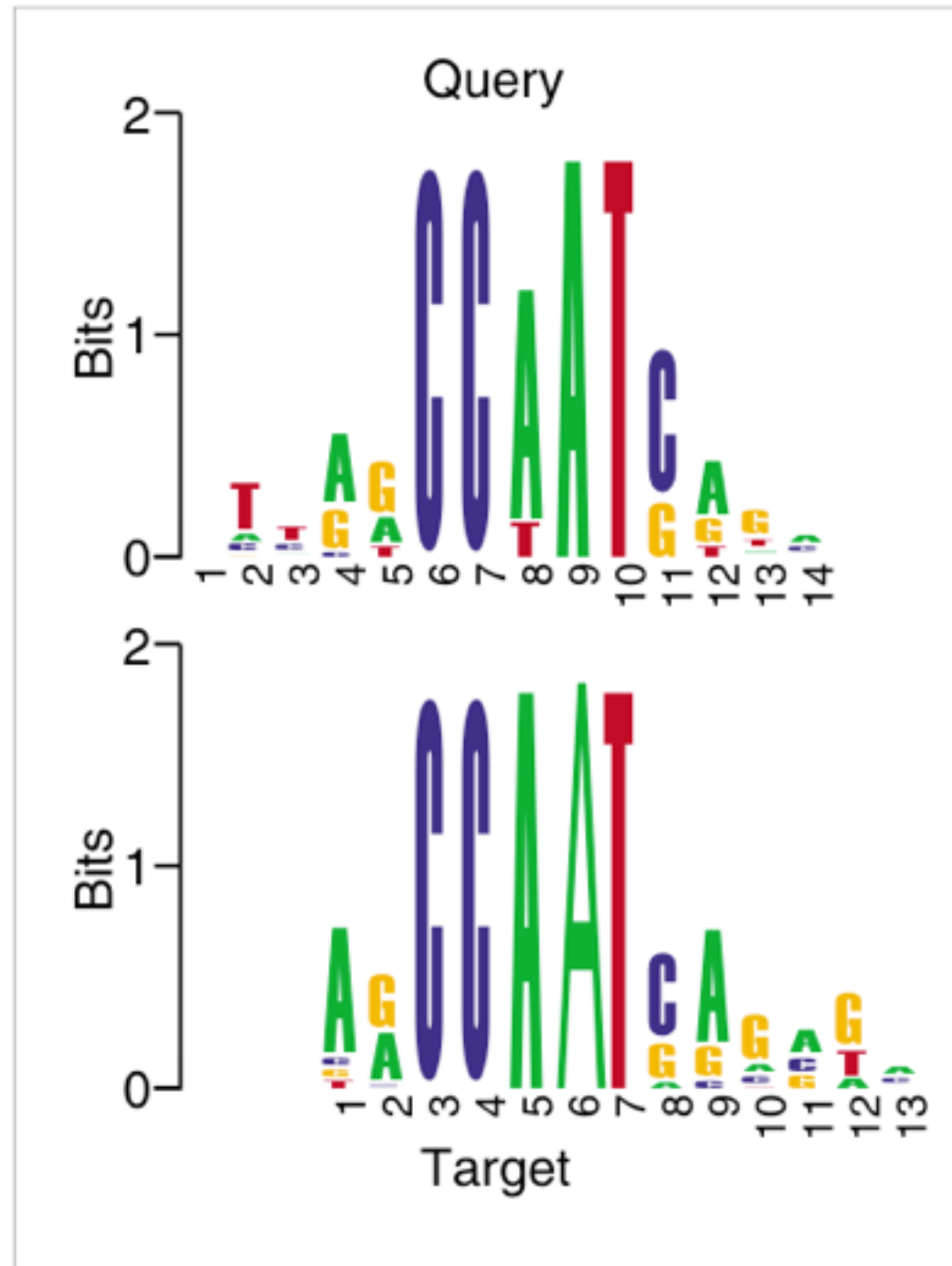


CAP-binding motif model
based on 59 binding sites in
E.coli

helix-turn-helix motif model
based on 100 aligned protein
sequences

Figure from Crooks et al., *Genome Research* 14:1188-90, 2004.

Motifs Logo



How to detect Motifs?

➡ Regular expression

These are derived from single conserved regions, which are reduced to consensus expressions for db searches

- they are *minimal expressions*, so sequence information is lost
- the more divergent the sequences used, the more fuzzy & poorly discriminating the pattern becomes

Alignment

GAVDFIALCDRYF

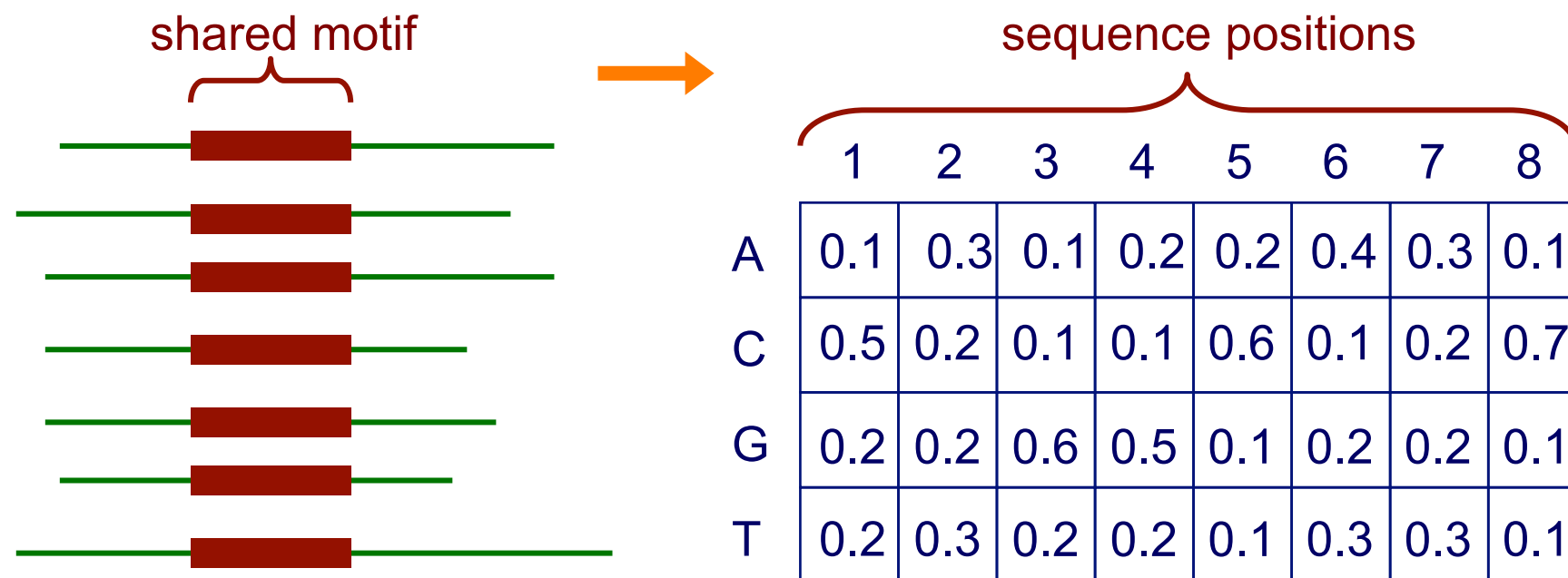
GPIDFVCF CERFY

GRVEFLNRCDRY

G-X-[IV]-[DE]-F-[IVL]-X2-C-[DE]-R-[FY]2

Motifs and *Position Weight Matrices*

- ➡ Given a set of aligned sequences, it is straightforward to construct a profile matrix characterizing a motif of interest



Each element represents the probability of given character at a specified position

Motifs and *Position Weight Matrices*

Motif Logo

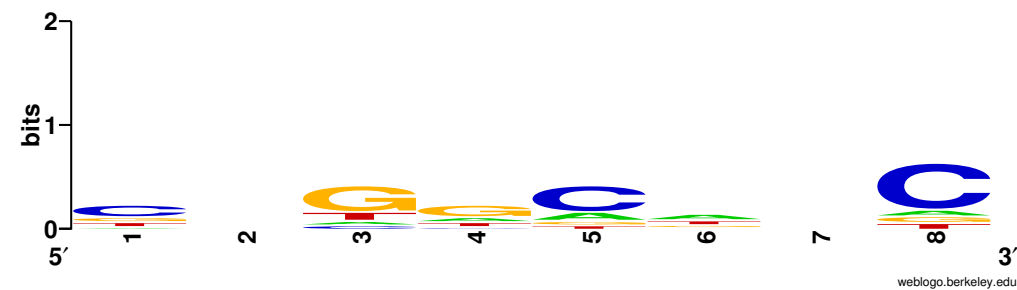
	1	2	3	4	5	6	7	8
A	0.1	0.3	0.1	0.2	0.2	0.4	0.3	0.1
C	0.5	0.2	0.1	0.1	0.6	0.1	0.2	0.7
G	0.2	0.2	0.6	0.5	0.1	0.2	0.2	0.1
T	0.2	0.3	0.2	0.2	0.1	0.3	0.3	0.1



or



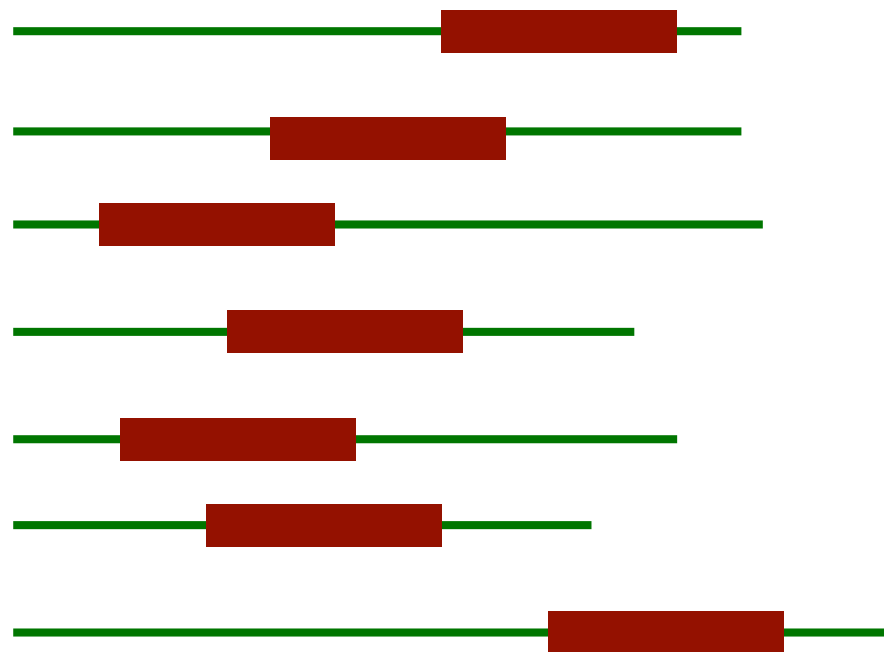
frequency logo



information content logo

Motifs and *Position Weight Matrices*

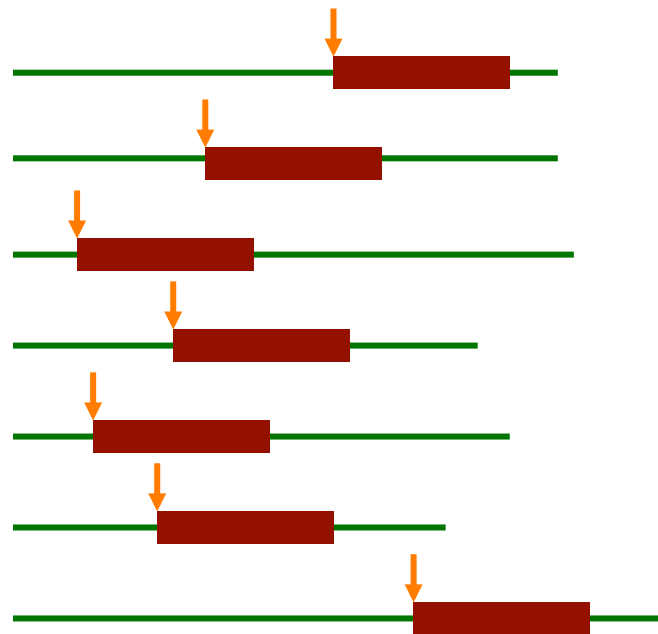
- ➡ How can we construct the matrix if the sequences aren't aligned?
- ➡ In the typical case we don't know what the motif looks like.



The Expectation-Maximization (EM) Approach

[Lawrence & Reilly, 1990; Bailey & Elkan, 1993, 1994, 1995]

- ➡ EM is a family of algorithms for learning probabilistic models in problems that involve *hidden state*
- ➡ In our problem, the hidden state is where the motif starts in each training sequence



Representing Motifs

A motif is

- assumed to have a fixed width, W
- represented by a matrix of probabilities: $p_{c,k}$ represents the probability of character c in column k

Also represent the “background”(i.e. sequence outside the motif): $p_{c,0}$ represents the probability of character c in the background

Representing Motifs

Example: a motif model of length 3

		0	1	2	3
p =	A	0.25	0.1	0.5	0.2
	C	0.25	0.4	0.2	0.1
	G	0.25	0.3	0.1	0.6
	T	0.25	0.2	0.2	0.1

background motif positions

Representing Motifs

➡ Suppose we are provided with label information that representing Motif Starting Positions

Example: given DNA sequences of length 6, where $W=3$

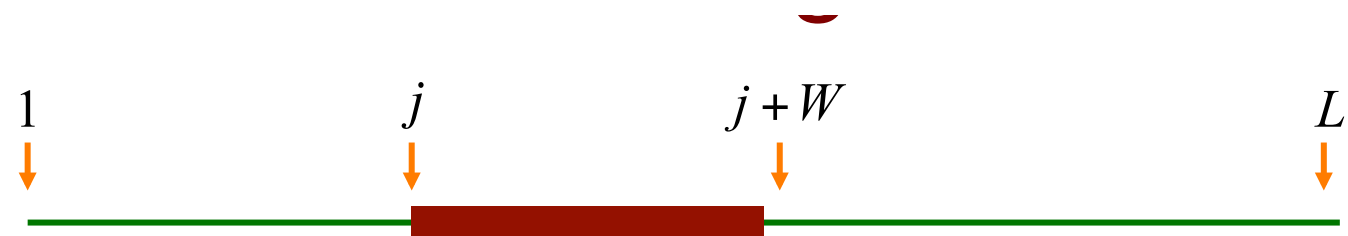
G T C A G G
G A G A G T
A C G G A G
C C A G T C

		1	2	3	4
$Z =$	seq1	0	0	1	0
	seq2	1	0	0	0
	seq3	0	0	0	1
	seq4	0	1	0	0

The element $Z_{i,j}$ of the matrix Z is an indicator random variable that takes value 1 if the motif starts in position j in sequence i (and takes value 0 otherwise)

Probability of a Sequence Given a Motif Starting Position

➡ Suppose we are provided with label information that representing Motif Starting Positions



$$P(X_i \mid Z_{i,j} = 1, p) = \underbrace{\prod_{k=1}^{j-1} p_{c_k, 0}}_{\text{before motif}} \underbrace{\prod_{k=j}^{j+W-1} p_{c_k, k-j+1}}_{\text{motif}} \underbrace{\prod_{k=j+W}^L p_{c_k, 0}}_{\text{after motif}}$$

X_i is the i th sequence

$Z_{i,j}$ is 1 if motif starts at position j in sequence i

c_k is the character at position k in sequence i

Probability of a Sequence Given a Motif Starting Position : Example

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

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

$$P(X_i \mid Z_{i3} = 1, p) =$$

$$p_{G,0} \times p_{C,0} \times p_{T,1} \times p_{G,2} \times p_{T,3} \times p_{A,0} \times p_{G,0} =$$

$$0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.1 \times 0.25 \times 0.25$$

Likelihood

$$\begin{aligned}P(D \mid p) &= \prod_i P(X_i \mid p) \\&= \prod_i \sum_j P(X_i \mid Z_{ij} = 1, p) P(Z_{ij} = 1) \\&= (L - W + 1)^{-n} \prod_i \sum_j P(X_i \mid Z_{ij} = 1, p)\end{aligned}$$

Parameter estimation

- ➡ Now suppose we do not know p . How to estimate it from the observed sequence data S_1, S_2, \dots, S_n ?
- ➡ One solution: calculate the likelihood of observing the provided n sequences for different values of p ,
- ➡ Pick the one with the largest likelihood, that is, to find p^* that

Basic EM Approach

given: length parameter W , training set of sequences

$t=0$

set initial values for $p^{(0)}$

do

$++t$

 re-estimate $Z^{(t)}$ from $p^{(t-1)}$ (E-step)

 re-estimate $p^{(t)}$ from $Z^{(t)}$ (M-step)

 until change in $p^{(t)} < \varepsilon$ (or change in likelihood is $< \varepsilon$)

return: $p^{(t)}, Z^{(t)}$

Example: Computing $Z^{(t)}$ from $p^{(t-1)}$

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

$$p^{(t-1)} = \begin{array}{c} \begin{array}{ccccc} & 0 & 1 & 2 & 3 \\ \text{A} & 0.25 & 0.1 & 0.5 & 0.2 \\ \text{C} & 0.25 & 0.4 & 0.2 & 0.1 \\ \text{G} & 0.25 & 0.3 & 0.1 & 0.6 \\ \text{T} & 0.25 & 0.2 & 0.2 & 0.1 \end{array} \end{array}$$

$$Z_{i,1}^{(t)} \propto P(X_i | Z_{i,1} = 1, p^{(t-1)}) = 0.3 \times 0.2 \times 0.1 \times 0.25 \times 0.25 \times 0.25 \times 0.25$$

$$Z_{i,2}^{(t)} \propto P(X_i | Z_{i,2} = 1, p^{(t-1)}) = 0.25 \times 0.4 \times 0.2 \times 0.6 \times 0.25 \times 0.25 \times 0.25$$

⋮

- then normalize so that

$$\sum_{j=1}^{L-W+1} Z_{i,j}^{(t)} = 1$$

Example: Computing $p^{(t)}$ from $z^{(t)}$

- recall $p_{c,k}$ represents the probability of character c in position k ; values for $k=0$ represent the background

$$p_{c,k}^{(t)} = \frac{n_{c,k} + d_{c,k}}{\sum_b (n_{b,k} + d_{b,k})}$$

pseudo-counts

$$n_{c,k} = \begin{cases} \sum_i \sum_{\{j | X_{i,j+k-1} = c\}} z_{i,j}^{(t)} & k > 0 \\ n_c - \sum_{j=1}^W n_{c,j} & k = 0 \end{cases}$$

sum over positions where c appears

total # of c 's in data set $\rightarrow n_c$

Example: Computing $p^{(t)}$ from $z^{(t)}$

A C A G C A

$$Z_{1,1}^{(t)} = 0.1, \quad Z_{1,2}^{(t)} = 0.7, \quad Z_{1,3}^{(t)} = 0.1, \quad Z_{1,4}^{(t)} = 0.1$$

A G G C A G

$$Z_{2,1}^{(t)} = 0.4, \quad Z_{2,2}^{(t)} = 0.1, \quad Z_{2,3}^{(t)} = 0.1, \quad Z_{2,4}^{(t)} = 0.4$$

T C A G T C

$$Z_{3,1}^{(t)} = 0.2, \quad Z_{3,2}^{(t)} = 0.6, \quad Z_{3,3}^{(t)} = 0.1, \quad Z_{3,4}^{(t)} = 0.1$$

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

$$p_{C,2}^{(t)} = \frac{Z_{1,1}^{(t)} + Z_{1,4}^{(t)} + Z_{2,3}^{(t)} + Z_{3,1}^{(t)} + 1}{Z_{1,1}^{(t)} + Z_{1,2}^{(t)} \dots + Z_{3,3}^{(t)} + Z_{3,4}^{(t)} + 4}$$

⋮

Exercices

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 weak (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 :

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 weak (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) ?