# Learning Sequence Motif Models Using Expectation Maximization (EM)

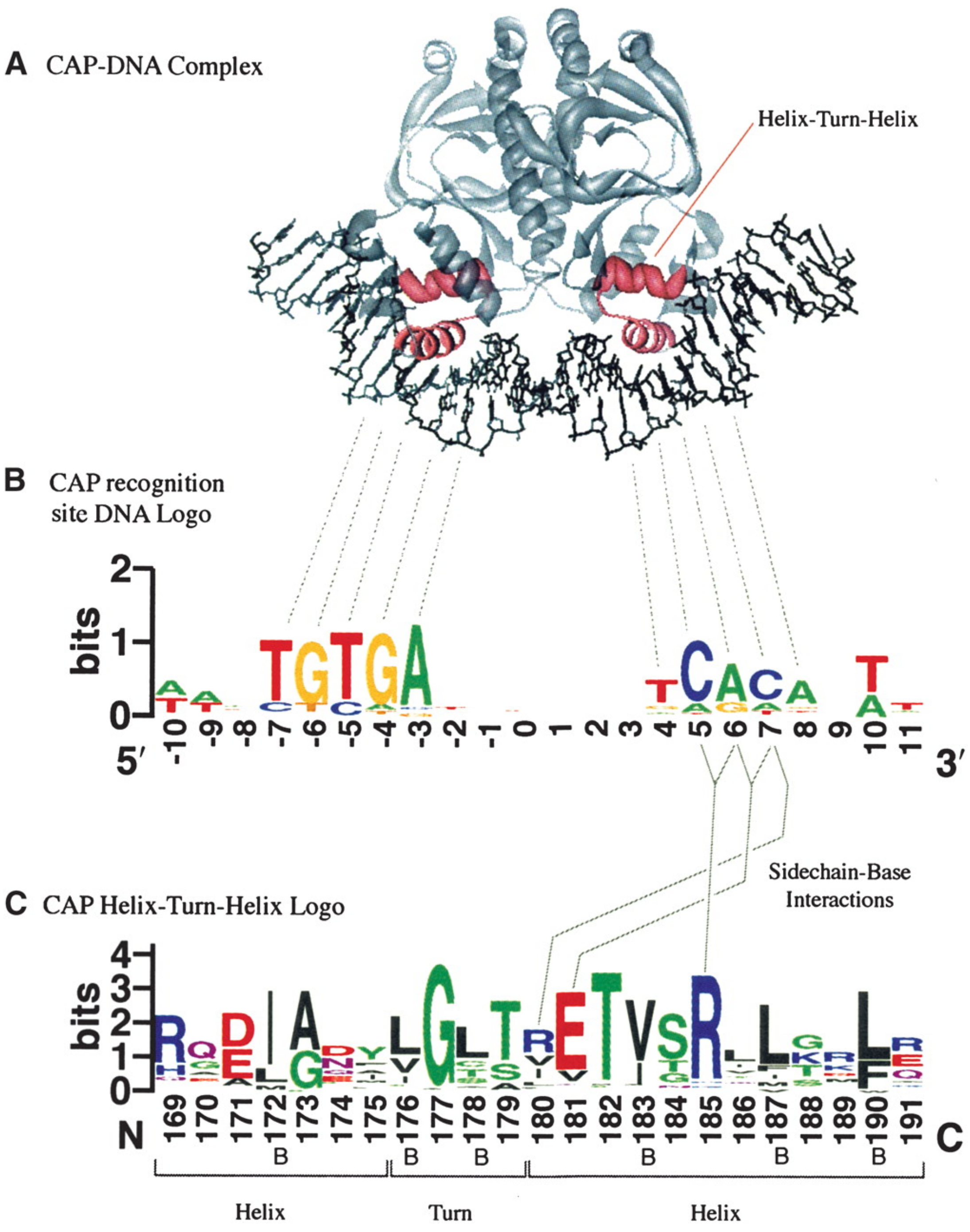
Juliana Silva Bernardes

# 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

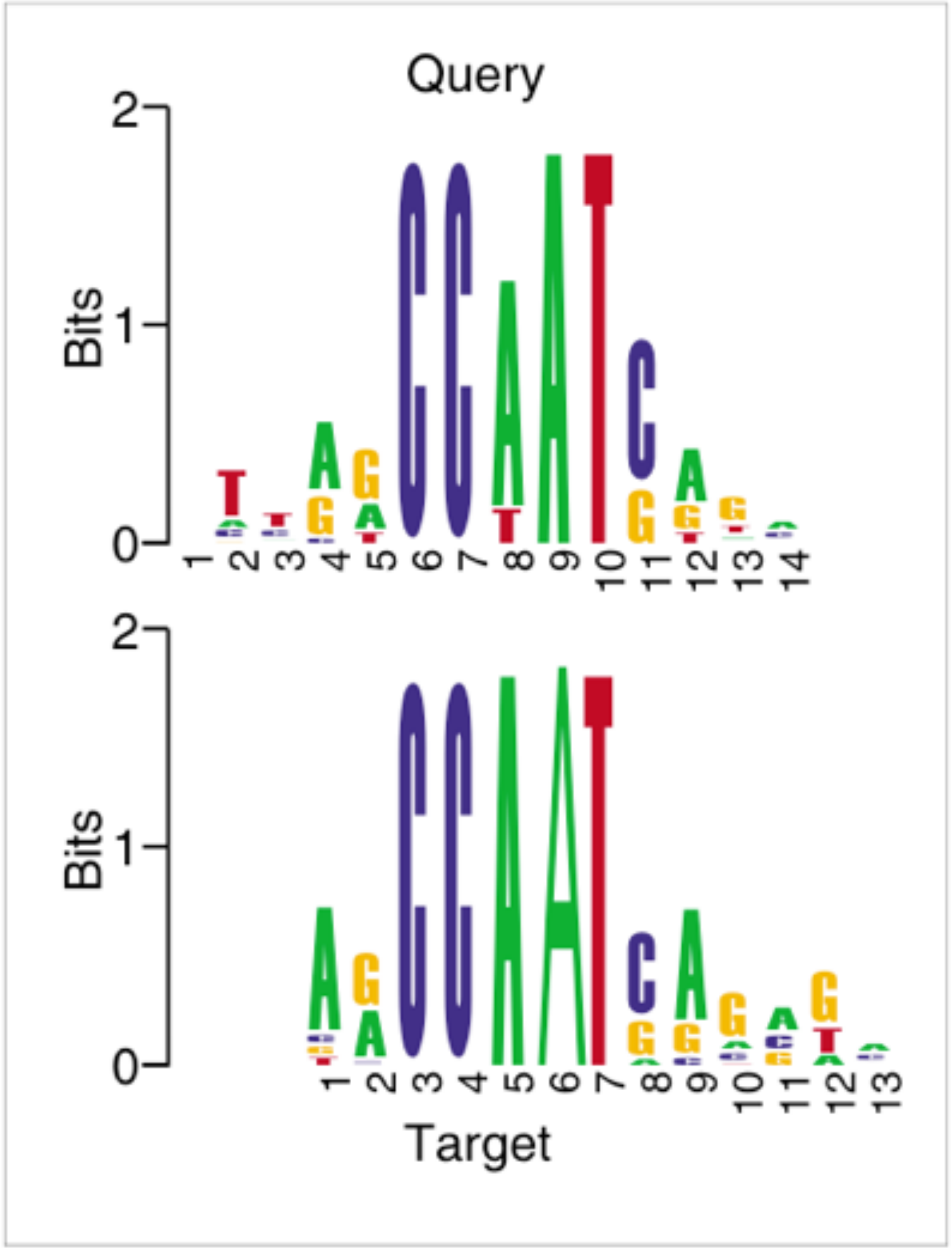


CAP-binding motif model based on 59 binding sites in E.coli

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

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

# Motifs Logo


Query
Bits
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Bits
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13
Target

# 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**

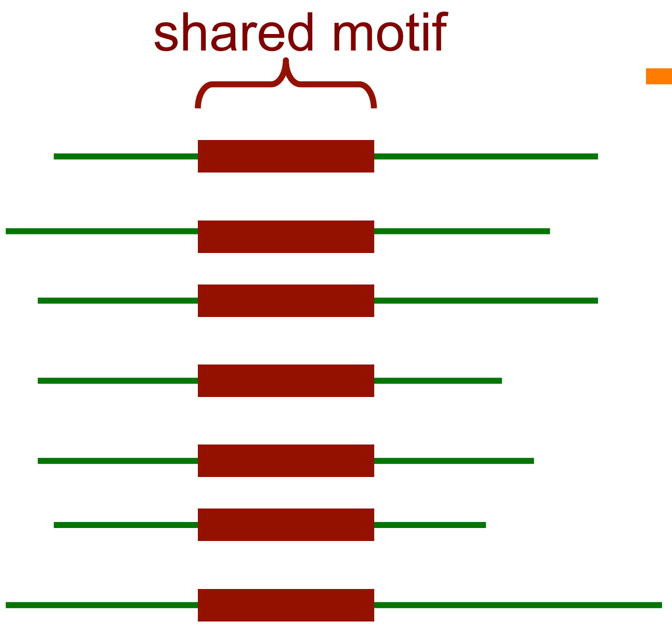
**Alignment**

```
GAVDFIALCDRYF
GPIDFVCFCERFY
GRVEFLNRCDRYY
```

**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



sequence positions

| | 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 |

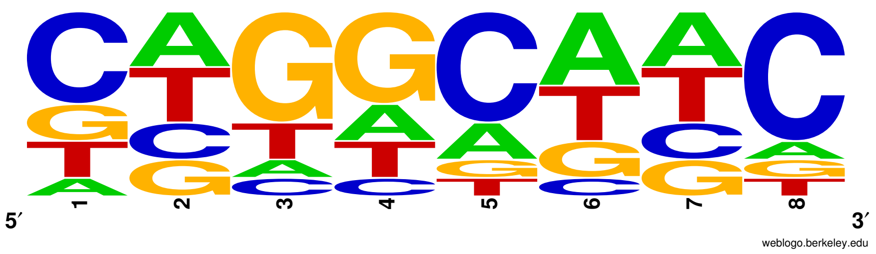
Each element represents the probability of given character at a specified position

# Motifs and *Position Weight Matrices*

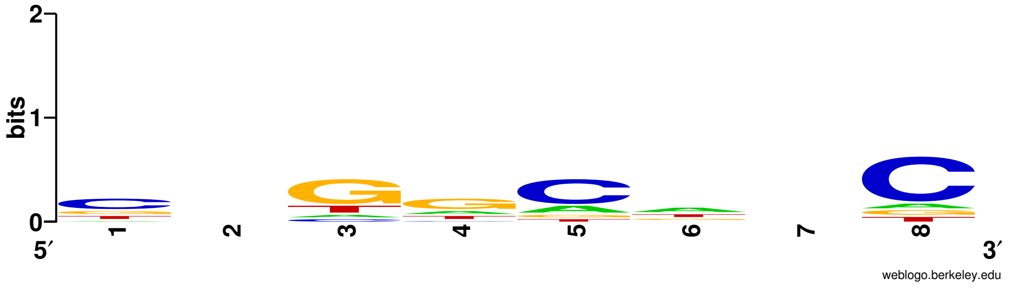
## Motif Logo

| | 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



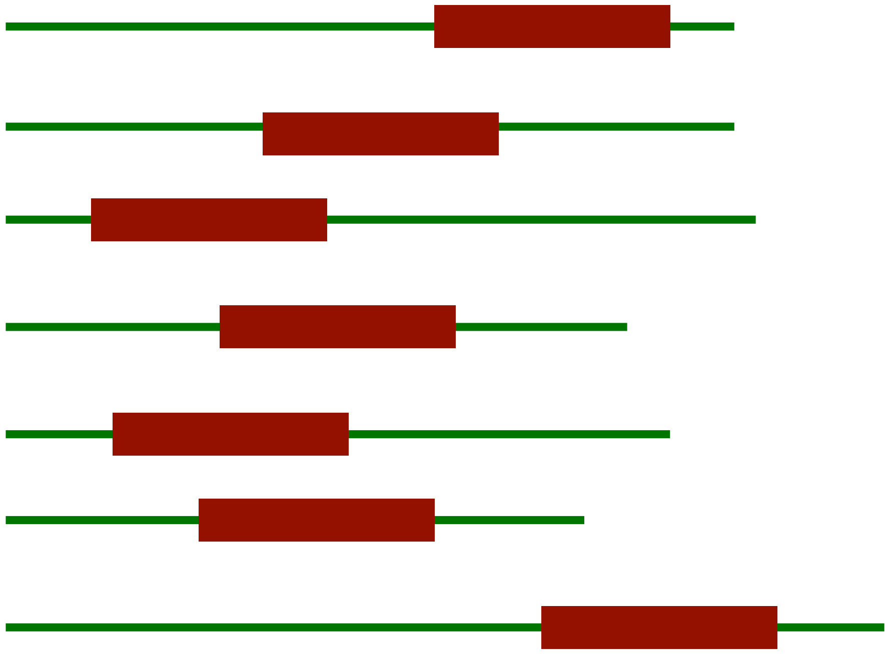
information content logo

# 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 :

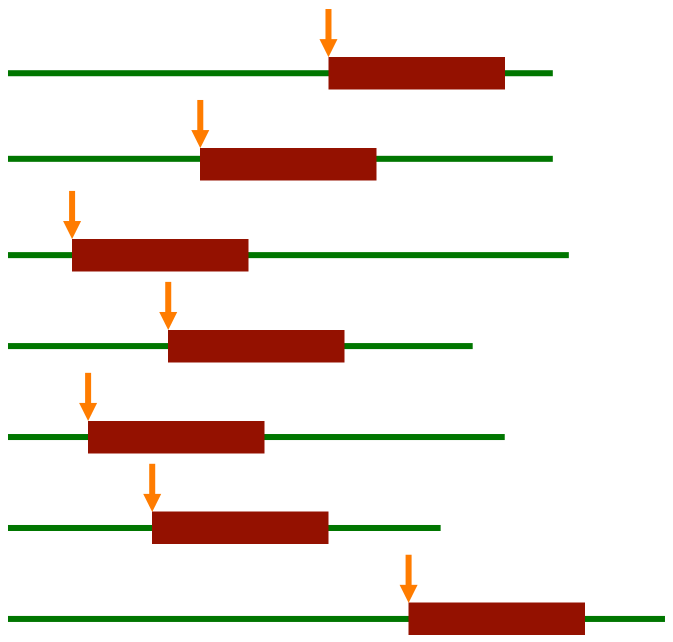
- what the motif looks like.
- where the motif starts
- How long the motif is?



# 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

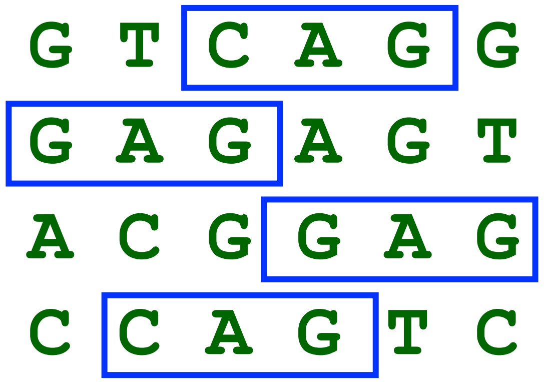
Example: a motif model of length 3

$$p = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \underbrace{\begin{array}{c} 0 \\ 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{array}}_{\text{background}} \quad \underbrace{\begin{array}{ccc} 1 & 2 & 3 \\ 0.1 & 0.5 & 0.2 \\ 0.4 & 0.2 & 0.1 \\ 0.3 & 0.1 & 0.6 \\ 0.2 & 0.2 & 0.1 \end{array}}_{\text{motif positions}}$$

# Representing Motifs

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

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



$$z = \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \text{seq1} & 0 & 0 & 1 & 0 \\ \text{seq2} & 1 & 0 & 0 & 0 \\ \text{seq3} & 0 & 0 & 0 & 1 \\ \text{seq4} & 0 & 1 & 0 & 0 \end{array}$$

The element $z_{i,j}$ of the matrix z is a 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
- What is the probability of observing a motif starting in position j of a given sequence Xi

$$P(X_i \mid z_{i,j} = 1, p) = \ ?$$

-Where z is a random variable that takes value 1 if the motif starts in position $j$ in sequence X$i$ (and takes value 0 otherwise)

-p is the motif model.

# Probability of a Sequence Given a Motif Starting Position

- ➡What is the probability of observing a motif starting in position j of a given sequence Xi

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

$X_i$ 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 $i$

# Example

- ➡What is the probability of observing a motif starting in position j of a given sequence Xi

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

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

$$Z\approx \quad P(X_i \mid z_{i3} = 1, p) =$$

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

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

# Likelihood

$$P(D \mid p) = \prod_i P(X_i \mid p)$$

$$= \prod_i \sum_j P(X_i \mid \mathrm{z_{ij}} = 1, \mathrm{p})P(\mathrm{z_{ij}} = 1)$$

$$= \left(L - W + 1\right)^{-n} \prod_i \sum_j P(X_i \mid \mathrm{z_{ij}} = 1, p)$$

# How to learn p?

**Parameter estimation :** Suppose we do not know p. How to estimate it from the observed sequence data $D = \{S_1, S_2, \cdots, S_n\}$ ?

➡One solution: calculate the likelihood of observing the provided $n$ sequences for different values of $p$,

➡Pick the one with the largest likelihood in order to find $p^*$

# 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 $p^{(t-1)}$

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

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

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

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

⋮

- 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

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

sum over positions where $c$ appears

total # of $c$'s in data set

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

**k>0**

X1= **A C A G C A**

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

X2= **A G G C A G**

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

X3= **T C A G T C**

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

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

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

$$\vdots$$

# Example: Computing $p_0^{(t)}$ from $z^{(t)}$

**k=0**

$$n_{A,0} = n_A - [n_{A,1} + n_{A,2} + n_{A,3}]$$

$n_A$=Total of A's within sequences X1, X2 and X3 = 6

$n_{A,1} = Z_{1,1} + Z_{1,3} + Z_{2,1} + Z_{3,3} \quad =0.7$

$n_{A,2} = Z_{1,2} + Z_{2,4} + Z_{3,2} \qquad =1.7$

$n_{A,3} = Z_{1,1} + Z_{1,4} + Z_{2,3} + Z_{3,1} \quad =0.5$

$$n_{A,0} = 6 - [0.7 + 1.7 + 0.5] =3.1$$

$$p^{(t)}{}_{A,0} = \frac{n_{A,0} + d_{A,0}}{n_{A,0} + \ldots + n_{G,0} + d_{A,0} + \ldots + d_{G,0}}$$

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

$$P(D \mid p) = \prod_i P(X_i \mid p)$$

$$= \prod_i \sum_j P(X_i \mid Z_{ij} = 1, p) P(Z_{ij} = 1)$$

$$= \left(L - W + 1\right)^{-n} \prod_i \sum_j P(X_i \mid Z_{ij} = 1, p)$$

n=3; L=6; W=3

$P(D|p) = (6\text{-}3+1)^{-3} *[P(X_1|\ z_{11}=1) + P(X_1|\ z_{12}=1) + P(X_1|\ Z_{13}=1) + P(X_1|\ z_{14}=1)]*$

$[P(X_2|\ z_{21}=1) + P(X_2|\ z_{22}=1) + P(X_2|\ z_{23}=1) + P(X_2|\ z_{24}=1)]*$

$[P(X_3|\ z_{31}=1) + P(X_3|\ z_{32}=1) + P(X_3|\ z_{33}=1) + P(X_3|\ z_{34}=1)]$

To compute $P(X_i|\ z_{ij}=1)$ see slide 15