Motif discovery

Jiří Kléma

Department of Computer Science,
Czech Technical University in Prague

Lecture based on Daifeng Wang's class at University of Wisconsin

IDA
Intelligent Data Analysis
RESEARCH GROUP

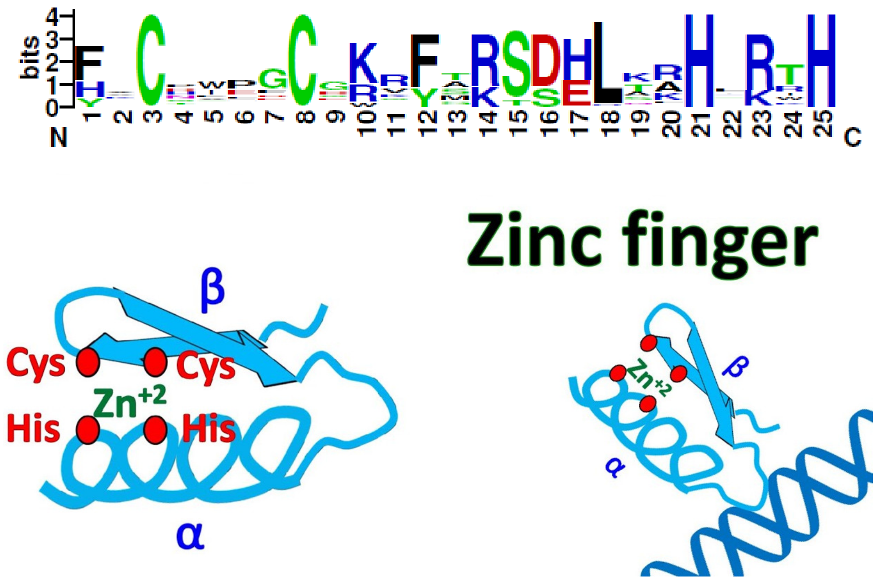
http://cw.felk.cvut.cz/wiki/courses/b4m36bin/start

Overview

- Sequence motifs
  - motivation, example,
  - definition,
  - (visual) representation,
- motif learning task
  - a solution with expectation maximization,
  - a solution with Gibbs sampling,
- untouched issues.

Sequence motif

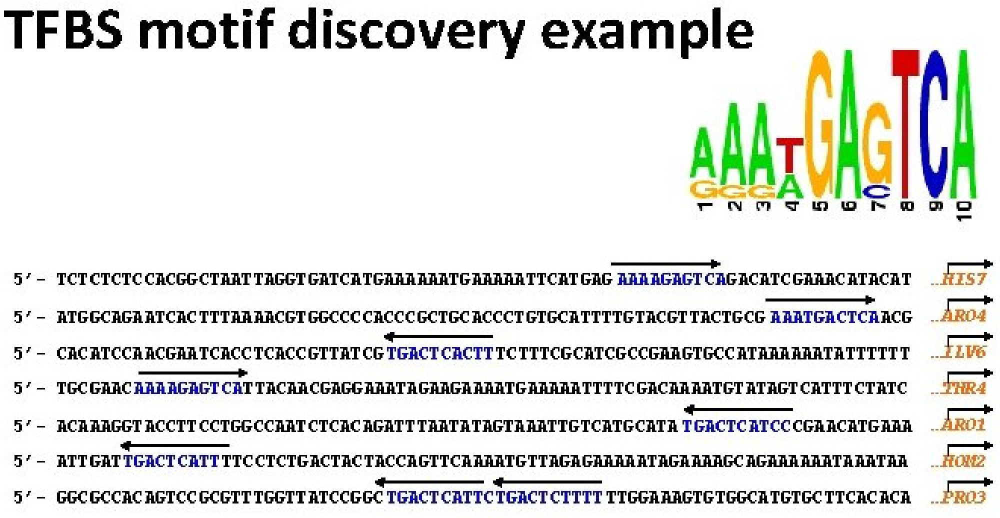
- A sequence motif
  - nucleotide or amino-acid **sequence pattern of biological significance**,
  - in the exon of a gene it may encode the “structural motif” of a protein.



Quick Biochemistry Basics.

Sequence motif

- A sequence motif
  - nucleotide or amino-acid **sequence pattern of biological significance**,
  - outside of gene exons, there exist regulatory sequence motifs, e.g., DNA sequences corresponding to protein binding sites, or motifs that control mRNA biogenesis or translation,
  - short coding motifs lack secondary structure and label proteins for delivery to particular parts of a cell, or mark them for phosphorylation.



TFBS motif discovery example

```
5′- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAG AAAAGAGTCAGACATCGAAACATACAT   ...HIS7
5′- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCG AAATGACTCAACG   ...ARO4
5′- CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT   ...ILV6
5′- TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC   ...THR4
5′- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATA TGACTCATCCCGAACATGAAA   ...ARO1
5′- ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA   ...HOM2
5′- GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTT TTGGAAAGTGTGGCATGTGCTTCACACA   ...PRO3
```

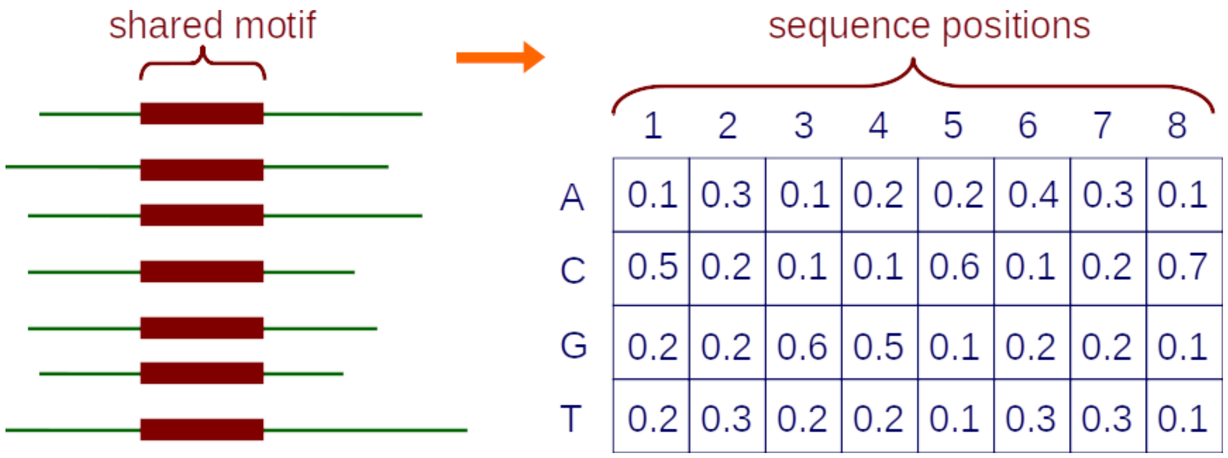
Canadian Bioinformatics Workshops.

Motif learning task

- Given:
  - a set of sequences that are thought to contain occurrences of an unknown motif of interest,
- Do:
  - infer a model of the motif
  - predict the locations of the motif occurrences in the given sequences.
- Why:
  - to understand which regions of sequences are functional, in particular:
    - DNA: mechanisms by which the expression of genes are regulated,
    - proteins: which regions interface with other molecules,
    - mutations in these regions may be significant (e.g., non-coding variants).

Sequence motif models

- Profile matrices (a.k.a. position weight matrices)
  - serve as probabilistic motif models,
  - other options: HMMs, regular expressions,
- 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 position.



	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

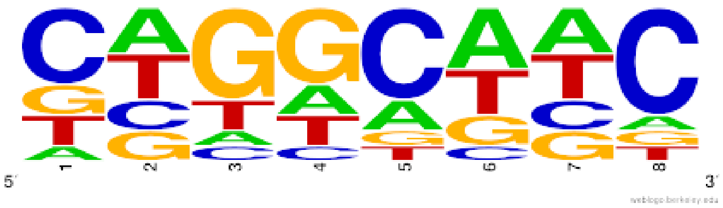
Wang: Learning Sequence Motif Models Using EM, Advanced Bioinformatics course.

Sequence logos

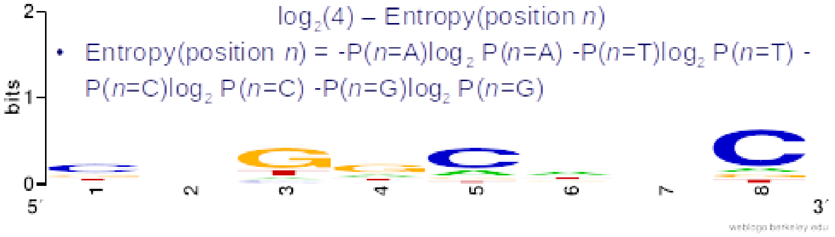
- Sequence logo is a graphical representation of profile matrices.

	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



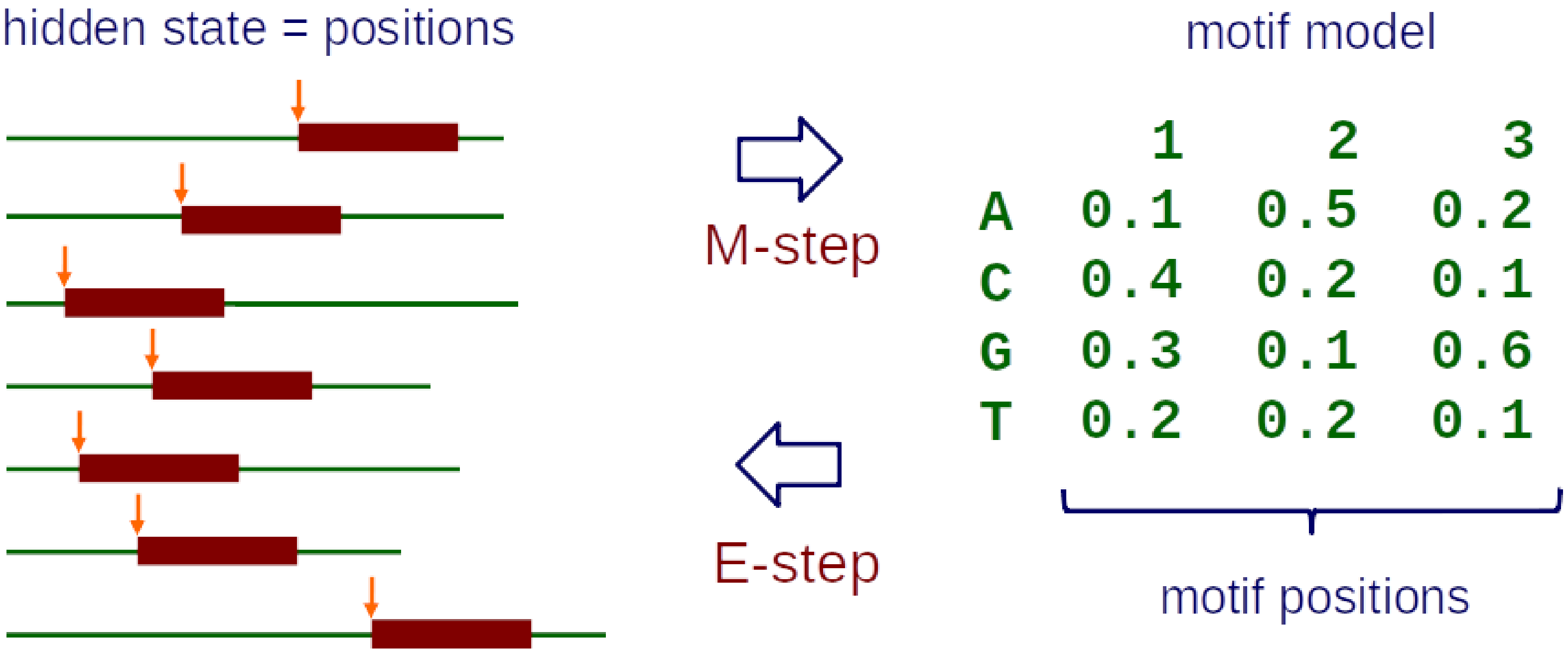
- Information content (IC) at position *n* = $\log_2(4)$ – Entropy(position *n*)
- Entropy(position *n*) = $-P(n{=}A)\log_2 P(n{=}A) - P(n{=}T)\log_2 P(n{=}T) - P(n{=}C)\log_2 P(n{=}C) - P(n{=}G)\log_2 P(n{=}G)$

information content logo

Wang: Learning Sequence Motif Models Using EM, Advanced Bioinformatics course.

Motifs and profile matrices in unaligned sequences

- As we do not know the motif we cannot know its positions/alignment too,
- there is a hidden state = where the motif starts in each training sequence,
- the task will have to be solved iteratively, e.g., with the EM algorithm.



motif model

	1	2	3
A	0.1	0.5	0.2
C	0.4	0.2	0.1
G	0.3	0.1	0.6
T	0.2	0.2	0.1

Wang: Learning Sequence Motif Models Using EM, Advanced Bioinformatics course.

Applying EM to the motif finding problem

- Identify the hidden variables Z
  - remember, they are the locations of the motifs,
- define the probabilistic model with parameters θ and likelihood function

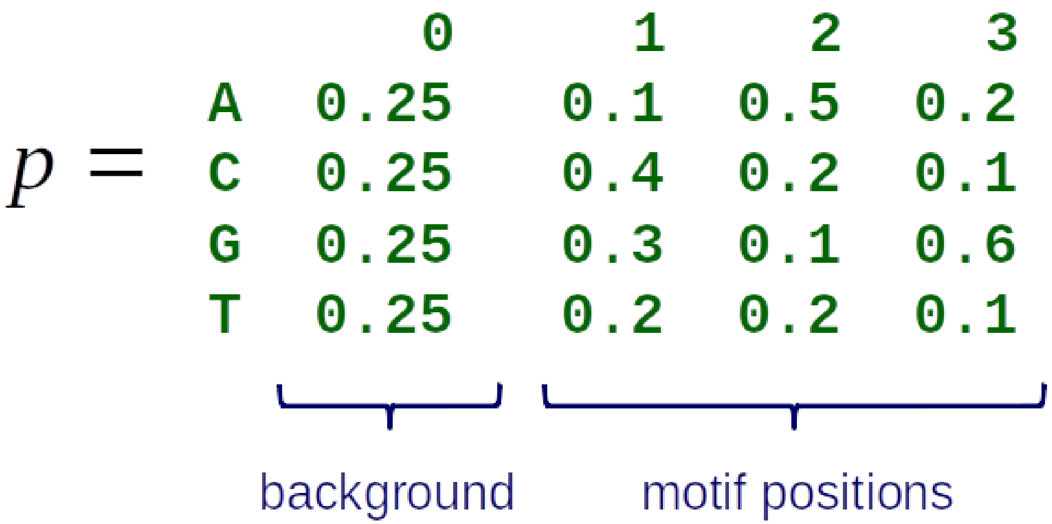
  $P(X|\theta) = \sum_Z P(X, Z|\theta)$,
  - where X stands for a set of sequences we learn from,
- write out the expectation (E) step
  - compute the expected values of the hidden variables given current parameter values θ^t,

    $Q(\theta|\theta^t) = \sum_Z P(Z|X, \theta^t) P(X, Z|\theta)$,
- write out the maximization (M) step
  - determine the parameters that maximize Q given the expected values of the hidden variables,

    $\theta^{t+1} = \arg\max_\theta Q(\theta|\theta^t)$.

Motif model (taken from MEME)

- MEME: Multiple EM for Motif Elicitation
  - 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 motif column k,
    - $p_{c,0}$ represent the background, i.e. sequence outside the motif,
  - example: a motif model of length 3 below.



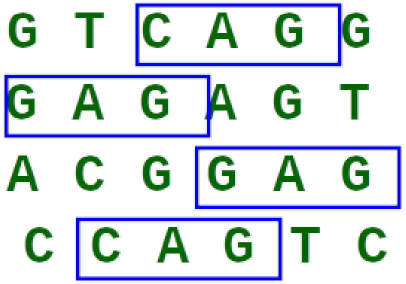
$$p = \begin{array}{ccccc} & 0 & 1 & 2 & 3 \\ A & 0.25 & 0.1 & 0.5 & 0.2 \\ C & 0.25 & 0.4 & 0.2 & 0.1 \\ G & 0.25 & 0.3 & 0.1 & 0.6 \\ T & 0.25 & 0.2 & 0.2 & 0.1 \end{array}$$

background
motif positions

Wang: Learning Sequence Motif Models Using EM, Advanced Bioinformatics course.

Motif starting positions (taken from MEME)

- MEME: Multiple EM for Motif Elicitation
  - a matrix Z, $Z_{i,j}$ takes value 1 if the motif starts in position j in sequence i (0 otherwise),
  - we will compute their expected values later,
  - example: given DNA sequences where $L = 6$ and $W = 3$, possible starting positions $m = L - W + 1$.

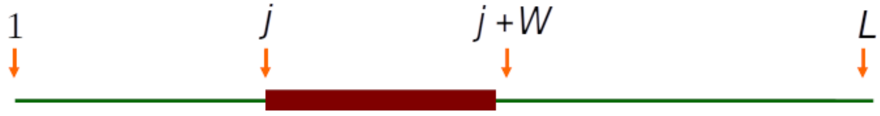


G T C A G G
G A G A G T
A C G G A G
C C A G T C

$$Z = \begin{array}{lcccc} & 1 & 2 & 3 & 4 \\ \text{seq1} & 0 & 0 & 1 & 0 \\ \text{seq2} & 1 & 0 & 0 & 0 \\ \text{seq3} & 0 & 0 & 0 & 1 \\ \text{seq4} & 0 & 1 & 0 & 0 \end{array}$$

Wang: Learning Sequence Motif Models Using EM, Advanced Bioinformatics course.

Probability of a sequence knowing motif starting position



$$P(X_i|Z_{i,j},p) = \prod_{k=1}^{j-1} p_{c_k,0} \prod_{k=j}^{j+W-1} p_{c_k,k-j+1} \prod_{k=j+W}^{L} p_{c_k,0}$$

- X_i is the i-th training sequence,
- $Z_{i,j}$ is 1 if motif starts at position j in sequence X_i,
- c_k is the character at position k in sequence X_i,

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

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

$$P(X_i|Z_{i,3}=1,p) =$$
$$= p_{G,0} \times p_{C,0} \times p_{T,1} \times p_{G,2} \times p_{T,3} \times p_{A,0} \times p_{G,0} =$$
$$= 0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.1 \times 0.25 \times 0.25$$

Basic EM approach

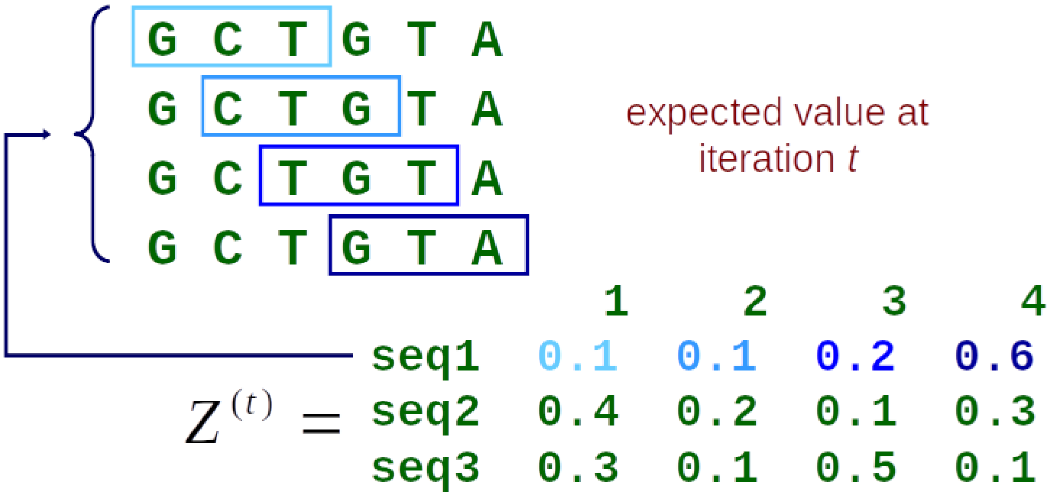
```
given: length parameter $W$,
       training set of sequences $X$

$t$=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 $p^{(t)} < \epsilon$ (or change in likelihood is $< \epsilon$)

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

The E-step: computing $Z^{(t)}$

- During the E-step, we compute the expected values of Z given X and $p^{(t-1)}$
  - $Z^{(t)} = E[Z|X, p^{(t-1)}]$,
  - where $Z^{(t)}$ stands for expected Z value at iteration t and Z for indicator random variable,



expected value at iteration *t*

$$Z^{(t)} = \begin{array}{lcccc} & 1 & 2 & 3 & 4 \\ \texttt{seq1} & 0.1 & 0.1 & 0.2 & 0.6 \\ \texttt{seq2} & 0.4 & 0.2 & 0.1 & 0.3 \\ \texttt{seq3} & 0.3 & 0.1 & 0.5 & 0.1 \end{array}$$

Wang: Learning Sequence Motif Models Using EM, Advanced Bioinformatics course.

The E-step: computing $Z^{(t)}$

- To estimate the starting positions in Z at step t we apply Bayes' rule to

$$P(Z_{i,j} = 1 | X_i, p^{(t-1)})$$

$$Z_{i,j}^{(t)} = \frac{P(X_i | Z_{i,j} = 1, p^{(t-1)}) P(Z_{i,j} = 1)}{\sum_{k=1}^{m} P(X_i | Z_{i,k} = 1, p^{(t-1)}) P(Z_{i,k} = 1)}$$

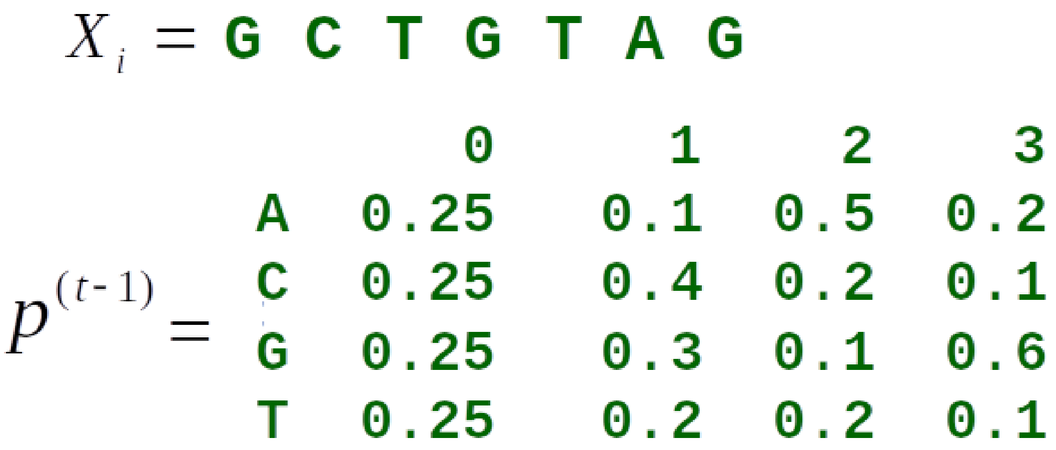
- if we assume that it is equally likely that the motif will start in any position

$$P(Z_{i,j} = 1) = \frac{1}{m}$$

$$Z_{i,j}^{(t)} = \frac{P(X_i | Z_{i,j} = 1, p^{(t-1)})}{\sum_{k=1}^{m} P(X_i | Z_{i,k} = 1, p^{(t-1)})}$$

The E-step: computing $Z^{(t)}$

- Let us show an example of $Z^{(t)}$ computation for one sequence



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

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

Wang: Learning Sequence Motif Models Using EM, Advanced Bioinformatics course.

$$Z_{i,1}^{(t)} \propto P(X_i|Z_{i,1}=1,p^{(t-1)}) = 0.3 \times 0.2 \times 0.1 \times 0.25 \times 0.25 \times 0.25 \times 0.25$$
$$Z_{i,2}^{(t)} \propto P(X_i|Z_{i,2}=1,p^{(t-1)}) = 0.25 \times 0.4 \times 0.2 \times 0.6 \times 0.25 \times 0.25 \times 0.25$$

- Eventually, normalize so that $\sum_{j=1}^{m} Z_{i,j}^{(t)} = 1$.

The M-step: estimating p

- Recall $p_{c,k}$ represents the probability of character c in k-th motif position
  - values for $k = 0$ represent the background,
- we will get them from observed values n and regularizing pseudocounts d
  - where n_c stands for the total number of cs in data,
  - and $n_{c,k}$ stands for the number of cs at position k.

$$p_{c,k}^{(t)} = \frac{n_{c,k} + d_{c,k}}{\sum_{b\in\{A,C,G,T\}}(n_{b,k} + d_{b,k})}$$

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

The M-step: estimating p

- Let us do a small example with 3 sequences:

ACAGCA $Z^{(t)}_{1,1} = 0.1, Z^{(t)}_{1,2} = 0.7, Z^{(t)}_{1,3} = 0.1, Z^{(t)}_{1,4} = 0.1$

AGGCAG $Z^{(t)}_{2,1} = 0.4, Z^{(t)}_{2,2} = 0.1, Z^{(t)}_{2,3} = 0.1, Z^{(t)}_{2,4} = 0.4$

TCAGTC $Z^{(t)}_{3,1} = 0.2, Z^{(t)}_{3,2} = 0.6, Z^{(t)}_{3,3} = 0.1, Z^{(t)}_{3,4} = 0.1$

$$p^{(t)}_{A,1} = \frac{Z^{(t)}_{1,1} + Z^{(t)}_{1,3} + Z^{(t)}_{2,1} + Z^{(t)}_{3,3} + 1}{Z^{(t)}_{1,1} + Z^{(t)}_{1,2} + \cdots + Z^{(t)}_{3,3} + Z^{(t)}_{3,4} + 4} = 0.24$$

$$p^{(t)}_{C,2} = \frac{Z^{(t)}_{1,1} + Z^{(t)}_{1,4} + Z^{(t)}_{2,3} + Z^{(t)}_{3,1} + 1}{Z^{(t)}_{1,1} + Z^{(t)}_{1,2} + \cdots + Z^{(t)}_{3,3} + Z^{(t)}_{3,4} + 4} = 0.21$$

What we have left untouched

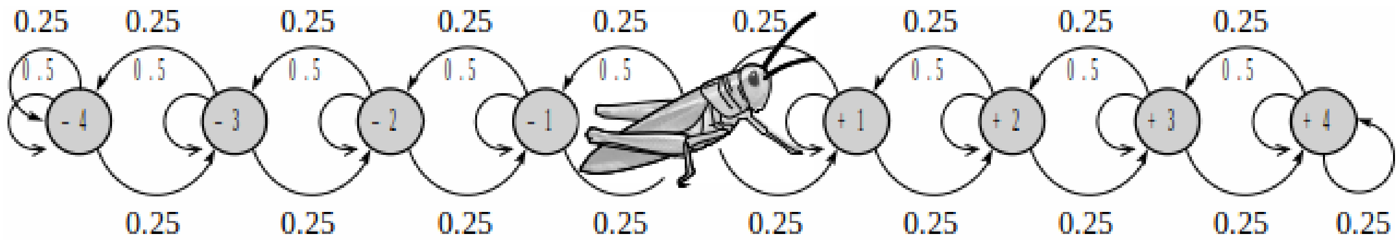
- We only solved OOPS (one motif occurrence per sequence)
  - this is not the general case,
  - ZOOPS (zero or one motif per sequence) is more general
    - EM includes another parameter γ for prior probability that a sequence contains a motif,
  - any number of repeats (ANR) is the most general approach,
- choosing the width of the motif,
- finding multiple motifs in a group of sequences,
- choosing good starting points for the parameters,
- using background knowledge to bias the parameters.

Gibbs Sampling: an alternative to EM

- EM can get trapped in local maxima
  - we may try different (perhaps random) initial parameters to alleviate this,
- Gibbs sampling exploits randomized search to a much greater degree
  - we can view it as stochastic analogy of EM for this task,
  - in theory, Gibbs sampling is less susceptible to local maxima than EM,
  - Gibbs will converge to a global maximum, in the limit,
  - probably not in a reasonable amount of time.
- in general, Gibbs sampling is a
  - Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution, when direct sampling is difficult.

Markov Chain Monte Carlo (MCMC) algorithms

- a Monte Carlo method
  - repeated random sampling serving to obtain numerical results,
- a Markov chain
  - a stochastic model of a sequence of events with limited memory,
- consider a Markov chain in which, on each time step, a grasshopper randomly chooses to stay in its current state, jump one state left or jump one state right



Koller and Friedman: Probabilistic Graphical Models, MIT Press.

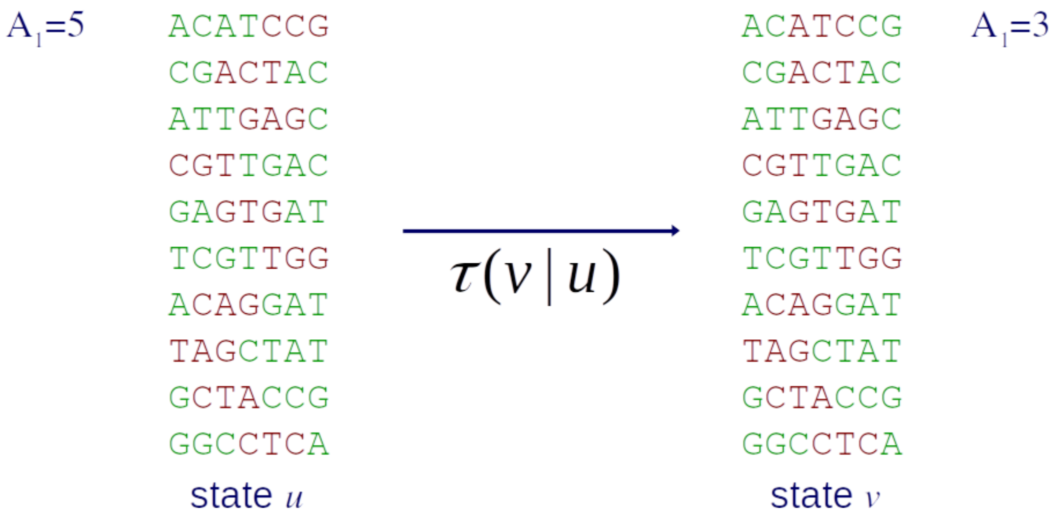
  - $P^{(t)}(u)$ is the probability of being in state u at time t in the random walk
    * $P^{(t+1)}(u) = \sum_v P^{(t)}(v)\tau(u|v)$, where τ is the transition probability,
    * $P^{(t+1)}(u) \approx P^{(t)}(u)$ for large t, becomes stationary.

MCMC with Gibbs sampling

- Gibbs sampling is a special case of MCMC in which
  - Markov chain transitions involve changing one variable at a time,
  - transition probability is conditional probability of the changed variable given all others,
  - we sample the joint distribution of a set of random variables $P(X_1, \dots, X_n)$ by iteratively sampling from $P(X_i | X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$.
- an example
  - Gibbs sampling for approximate inference in Bayesian networks,
  - the joint distribution is not directly available,
  - however, the network provides the conditional probabilities.

Gibbs sampling for motif learning

- In the EM approach we maintained a distribution $Z_i^{(t)}$ over the possible motif starting points for each sequence at iteration t,
- now, we will maintain a specific motif starting point a_i for each sequence, but we will keep randomly resampling them,
- Markov chain states will be the configurations of starting positions (a_i values for a set of random variables $\{A_1, \ldots, A_n\}$),
- transitions between states correspond to changing selected starting positions.



Wang: Learning Sequence Motif Models Using EM, Advanced Bioinformatics course.

Sampling with MCMC in general

- Want to find the mode of a certain distribution $\arg\max_x P(X)$,
- and it is intractable to do it directly,
- construct a Markov chain with
  - states corresponding to configurations of X,
  - stationary distribution equal to $P(X)$,
- through MCMC we can reconstruct the distribution and find the mode,
- the transition probabilities must keep the condition of **detailed balance**
  - $P(u)\tau(v|u) = P(v)\tau(u|v)$ for all pairs of states,
- then if we perform MCMC with N samples and $count(u)$ is the number of times we are in state u it holds that

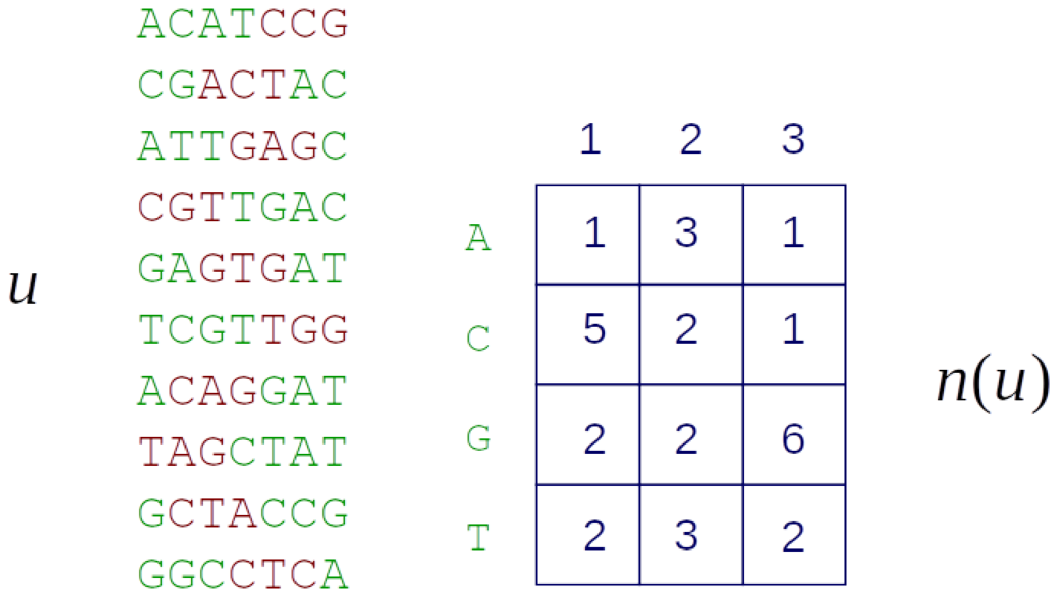
  $\frac{1}{N}\lim_{N\to\infty} count(u) = P(u)$.

Estimating the state probability and p

- The probability of a state is given by

$$P(u) \propto \prod_{c} \prod_{j=1}^{W} \left(\frac{p_{c,j}}{p_{c,0}}\right)^{n_{c,j}(u)}$$

- where $n_{c,j}(u)$ is the count of c in motif position j,
- $p_{c,j}$ is the probability of c in motif position j and $p_{c,0}$ its background probability.



u

ACATCCG
CGACTAC
ATTGAGC
CGTTGAC
GAGTGAT
TCGTTGG
ACAGGAT
TAGCTAT
GCTACCG
GGCCTCA

	1	2	3
A	1	3	1
C	5	2	1
G	2	2	6
T	2	3	2

$n(u)$

Wang: Learning Sequence Motif Models Using EM, Advanced Bioinformatics course.

Estimating the state probability and p

- Recall $p_{c,k}$ represents the probability of character c in k-th motif position, $k = 0$ represents the background

  EM:

$$p_{c,k}^{(t)} = \frac{n_{c,k} + d_{c,k}}{\sum_{b \in \{A,C,G,T\}} (n_{b,k} + d_{b,k})}$$

  Gibbs sampling:

$$p_{c,k}^{(t)} = \frac{n_{c,k} + d_c}{N - 1 + d_b}$$

$$p_{c,0} = \frac{n_{c,0} + d_c}{(N-1)(L-W) + d_b}$$

- where N is the number of sequences,
- L is the sequence length and W is motif length.

Sampling new motif positions

- For sampling a new motif position in sequence i,
- Estimate p from all sequences **except sequence i**,
- For each possible starting position $A_i = j$ compute the likelihood ratio

$$LR(j) = \frac{\prod_{k=j}^{j+W-1} p_{c_k,k-j+1}}{\prod_{k=j}^{j+W-1} p_{c_k,0}}$$

- Randomly select a new starting position $A_i = j$ with probability

$$\frac{LR(j)}{\sum_{k\in\{positions\}} LR(k)}$$

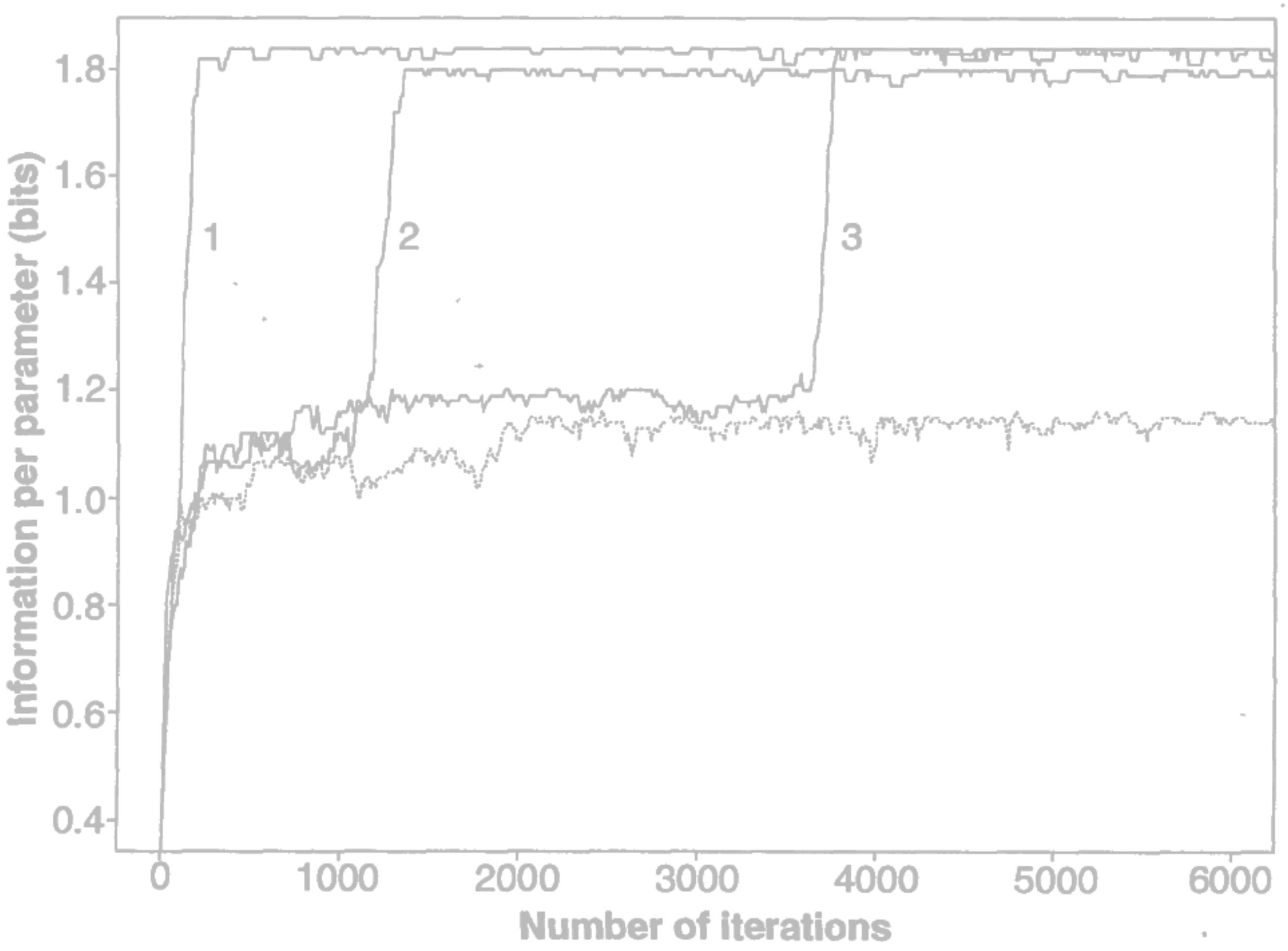
Gibbs sampling algorithm for motif finding

```
given: length parameter $W$
       training set of sequences

choose random positions for $a$
do
   pick a sequence $X_i$
   predictive update step:
    estimate $p$ given current motif positions $a$
    (using all sequences but $X_i$)
   sampling step:
    sample a new motif position $a_i$ for $X_i$
until convergence

return: $p$, $a$
```

Gibbs sampling: performance



Lawrence et al.: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment", Science.

Summary

- Motif discovery
  - local multiple alignments (compare with MSA discussed earlier),
- EM and Gibbs sampling discussed
  - many other methods exist,
  - including those that extract from MSA such as EMOTIF or PRINTS,
- in practice, motif finders often fail
  - motif signal could be too weak,
  - large search space with many local maxima,
- improvements through utilization of background knowledge
  - tying parameters,
  - (Dirichlet) priors.