Introducción a la Bioinformática:

Motif Discovery

Luis Garreta

Doctorado en Ingeniería
Pontificia Universidad Javeriana – Cali

March 15, 2017

Outline

What are DNA sequence Motifs?

Sequence motifs are short, recurring patterns in DNA that are presumed to have a biological function

What are motifs representing?

- Motifs represent a short common sequence
  - Regulatory motifs (Transcription Factors binding sites)
  - Functional site in proteins (DNA binding motif)

Regulatory Motifs

- Transcription Factors bind to regulatory motifs
  - Motifs are 6 – 20 nucleotides long
  - Activators and repressors
  - Usually located near target gene, mostly upstream
- Every gene contains a regulatory region (RR) typically stretching 100~1000 bp upstream of the transcriptional start site
- Located within the RR ar tge TFBS (motifs) specific for a given TF

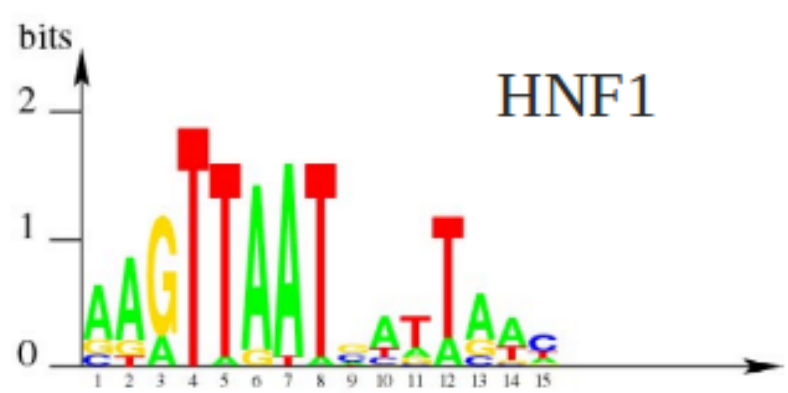
TFs influence gene expression by binding to a specific location in the respective gene's RR (TFBS)

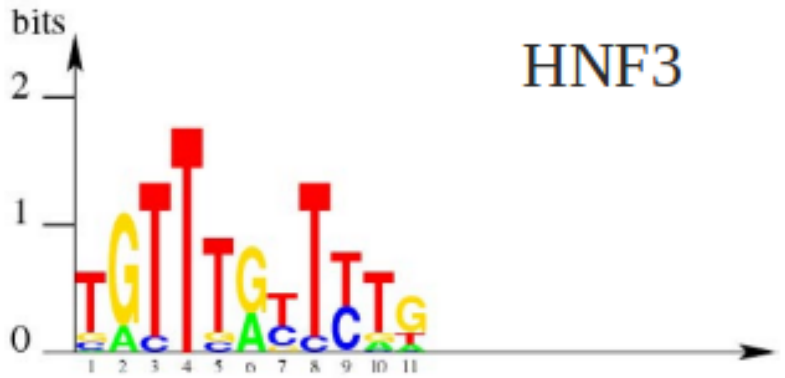
Motifs vs Transcription Factor Binding Sites

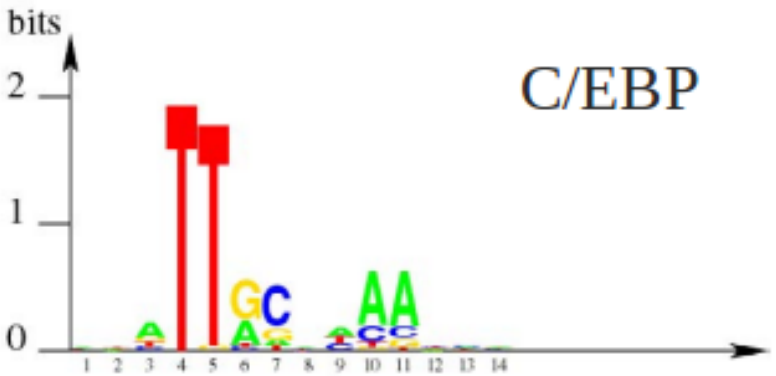
- Motifs:
  - Statistical or Computational entities
  - Predicted
- Transcription Factor Binding Sites (Cys-regulatory elements):
  - Biological entities
  - Real

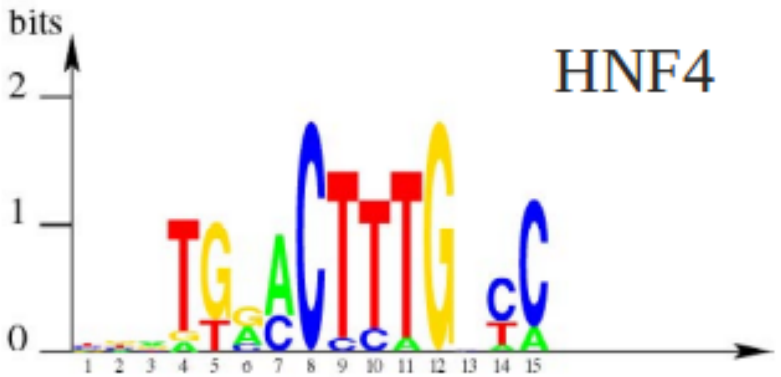
The hope is that TFBS are conserved (significant computationally), so motifs can be used to find them

Motif Logos for Liver TFBS









Motif Finding Problem (simple)

Given n sequences, find a motif (or subsequence) present in many

A String Search Problem: XXXXXXX

- Giving a random sample of DNA sequences:

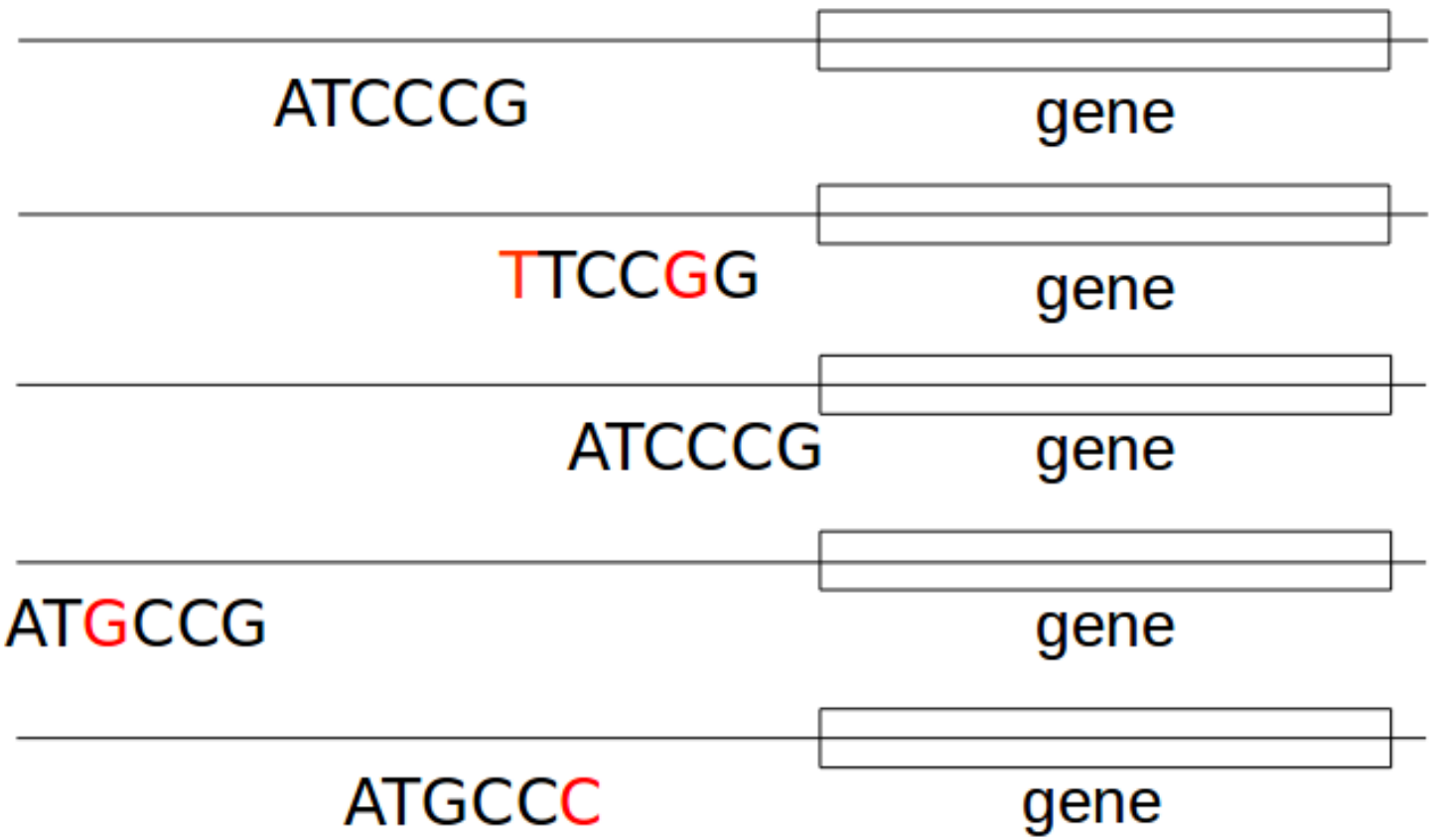
```
cctgatagacgctatctggctatccacgtacgtaggtcctctgtgcgaatctatgcgtttccaaccat
agtactggtgtacatttgatacgtacgtacaccggcaacctgaaacaaacgctcagaaccagaagtgc
aaacgtacgtgcaccctctttcttcgtggctctggccaacgagggctgatgtataagacgaaaatttt
agcctccgatgtaagtcatagctgtaactattacctgccacccctattacatcttacgtacgtataca
ctgttatacaacgcgtcatggcggggtatgcgttttggtcgtcgtacgctcgatcgttaacgtacgtc
```

- Find the patter that is implanted in each of the individual sequences, namely, the motif

A String Search Problem: acgtacgtX

- 5 sequences of 70 nucleotidas
- Motif size: 9 nucleotides (bp)

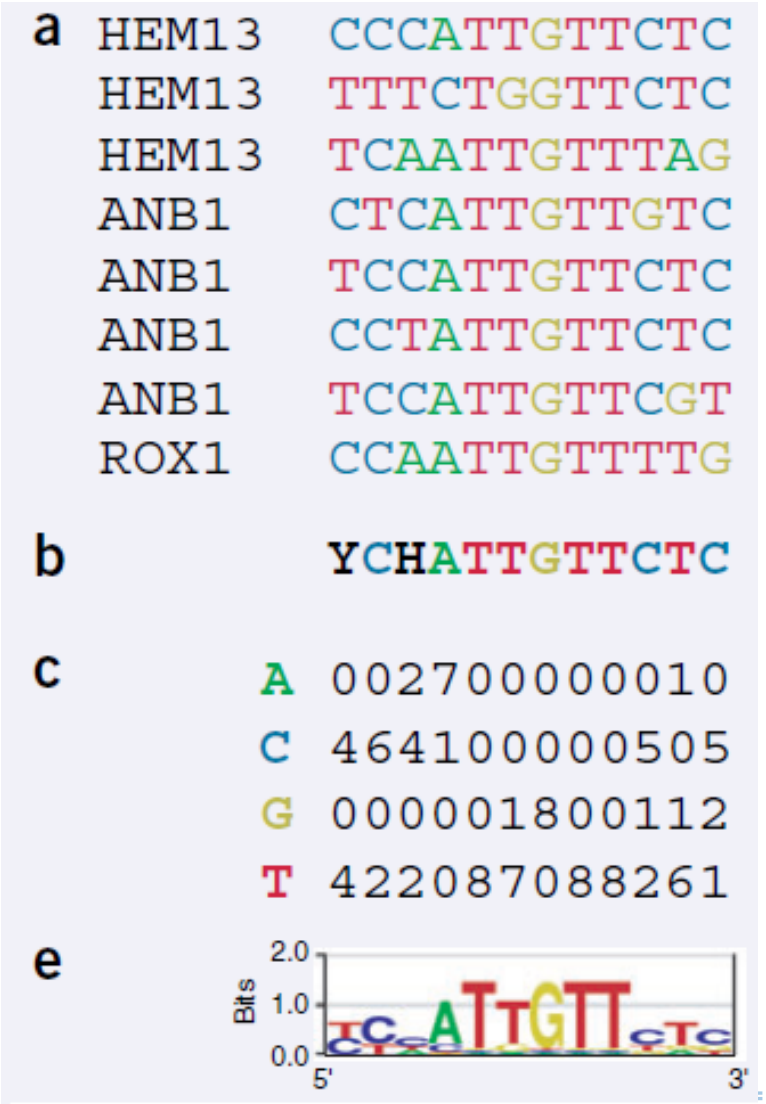
Motifs and Transcriptional Start Sites



Computational Motifs

Motifs Representation

- (a) Binding sites in three genes
- (b) Degenerate consensus sequence
- (c) Counts of nucleotides at each position
- (d) Sequence Logo showing Frequencies
- (e) Frequencies scaled relative to the information content



Position Weight Matrix (PWM)

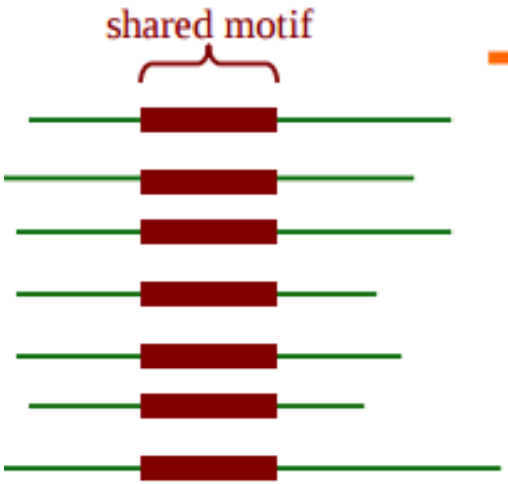
- Introduced as an alternative to consensus sequences
- Used to represent patterns in biological sequences
- Essential component in algorithms for motif discovery

PWMs or PSSMs or Profile Matrices

PSSM: Position-Specific Scoring Matrix

PWMs, PSSMs, Profile Matrices

Constructed from a set of aligned sequences and characterizing a motif of interest

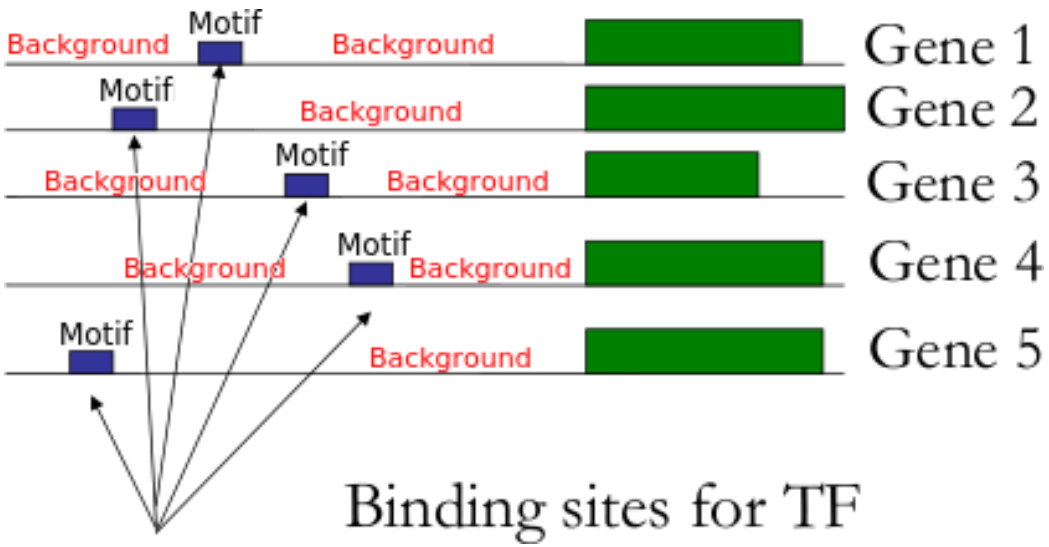


sequence positions

	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

Construction of a PWM:

From Binding Sites to a set of aligned sequences



GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT

Construction of a PWM:

From aligned sequences to a Position Count Matrix (PCM)

Formally, given a set X of N aligned sequences of length l, the elements of the PPM M are calculated:

X=

GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT

$$M_{k,j} = \sum_{i=1}^{N} I(X_{i,j} = k),$$

$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}$$

Construction of a PWM:

From a PCM to a PPM or PFM

- PPM: Position Probability Matrix
- PFM: Position Frequency Matrix[º]

$$M = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix} \qquad M_{k,j} = \sum_{i=1}^{N} I(X_{i,j} = k),$$

$$\Downarrow$$

$$M = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{bmatrix} .3 & .6 & .1 & .0 & .0 & .6 & .7 & .2 & .1 \\ .2 & .2 & .1 & .0 & .0 & .2 & .1 & .1 & .2 \\ .1 & .1 & .7 & .0 & .0 & .1 & .1 & .5 & .1 \\ .4 & .1 & .1 & .0 & .0 & .1 & .1 & .2 & .6 \end{bmatrix} \qquad M_{k,j} = \frac{1}{N} \sum_{i=1}^{N} I(X_{i,j} = \quad : k),$$

Construction of a PWM:

Calculating the probability of a sequence given the PPM

$$S = GAGGTAAAC$$

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} .3 & .6 & .1 & .0 & .0 & .6 & .7 & .2 & .1 \\ .2 & .2 & .1 & .0 & .0 & .2 & .1 & .1 & .2 \\ .1 & .1 & .7 & .0 & .0 & .1 & .1 & .5 & .1 \\ .4 & .1 & .1 & .0 & .0 & .1 & .1 & .2 & .6 \end{bmatrix}$$

Probability of S given M

$$P(S|M) = 0.1x0.6x0.7x1.0x1.0x0.6x0.7x0.2 = 0.0007056$$

Construction of a PWM:

Background Model

Represents the set of sequences having a frequency distribution different from the motif

The simplest background model assumes that each letter appears equally frequently in the dataset, that is:
$b_k = 1/|k|$

Background =

A=0.25
C=0.25
G=0.25
T=0.25

Construction of a PWM:

Calculating the PWM values with no pseudocounts added

PPM:

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} .3 & .6 & .1 & .0 & .0 & .6 & .7 & .2 & .1 \\ .2 & .2 & .1 & .0 & .0 & .2 & .1 & .1 & .2 \\ .1 & .1 & .7 & .0 & .0 & .1 & .1 & .5 & .1 \\ .4 & .1 & .1 & .0 & .0 & .1 & .1 & .2 & .6 \end{bmatrix} \quad M_{k,j} = \frac{1}{N} \sum_{i=1}^{N} I(X_{i,j} = \ : k),$$

$\Downarrow$

- From a PCM to PWM
  - $M'_{k,j} = M_{k,j}/b_k$, with $b_k = 1/4$

$$M' = \begin{array}{c} \\ A: \\ C: \\ G: \\ T: \end{array} \begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1.20 & 2.40 & 0.40 & 0.00 & 0.00 & 2.40 & 2.80 & 0.80 & 0.40 \\ 0.80 & 0.80 & 0.40 & 0.00 & 0.00 & 0.80 & 0.40 & 0.40 & 0.80 \\ 0.40 & 0.40 & 2.80 & 4.00 & 0.00 & 0.40 & 0.40 & 2.00 & 0.40 \\ 1.60 & 0.40 & 0.40 & 0.00 & 4.00 & 0.40 & 0.40 & 0.80 & 2.40 \end{bmatrix}$$

Construction of a PWM:

Calculating the PWM as log likelihoods

PPM:

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} .3 & .6 & .1 & .0 & .0 & .6 & .7 & .2 & .1 \\ .2 & .2 & .1 & .0 & .0 & .2 & .1 & .1 & .2 \\ .1 & .1 & .7 & .0 & .0 & .1 & .1 & .5 & .1 \\ .4 & .1 & .1 & .0 & .0 & .1 & .1 & .2 & .6 \end{bmatrix} \quad M_{k,j} = \frac{1}{N} \sum_{i=1}^{N} I(X_{i,j} = \ : k),$$

$\Downarrow$

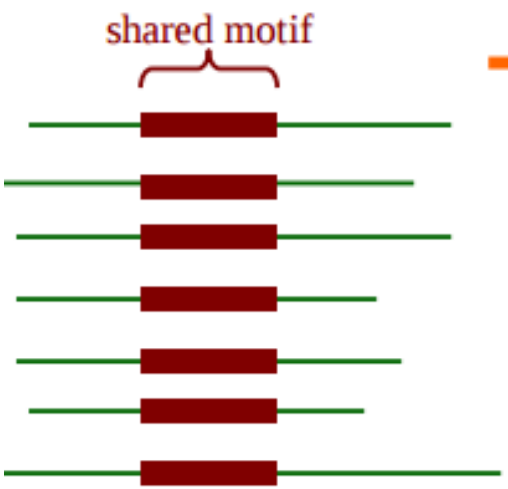
- From a PCM to PWM
  - $M'_{k,j} = M_{k,j}/b_k$, with $b_k = 1/4$

M'=

	0	1	2	3	4	5	6	7	8
A:	1.20	2.40	0.40	0.00	0.00	2.40	2.80	0.80	0.40
C:	0.80	0.80	0.40	0.00	0.00	0.80	0.40	0.40	0.80
G:	0.40	0.40	2.80	4.00	0.00	0.40	0.40	2.00	0.40
T:	1.60	0.40	0.40	0.00	4.00	0.40	0.40	0.80	2.40

The Problem

Motifs and PWMs: Known Motifs

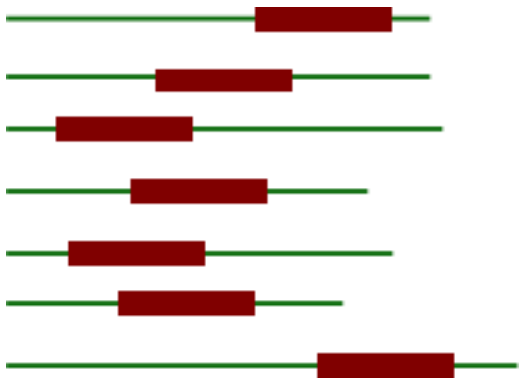


sequence positions

	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

Motifs and Profile Matrices: Unknown Motifs

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

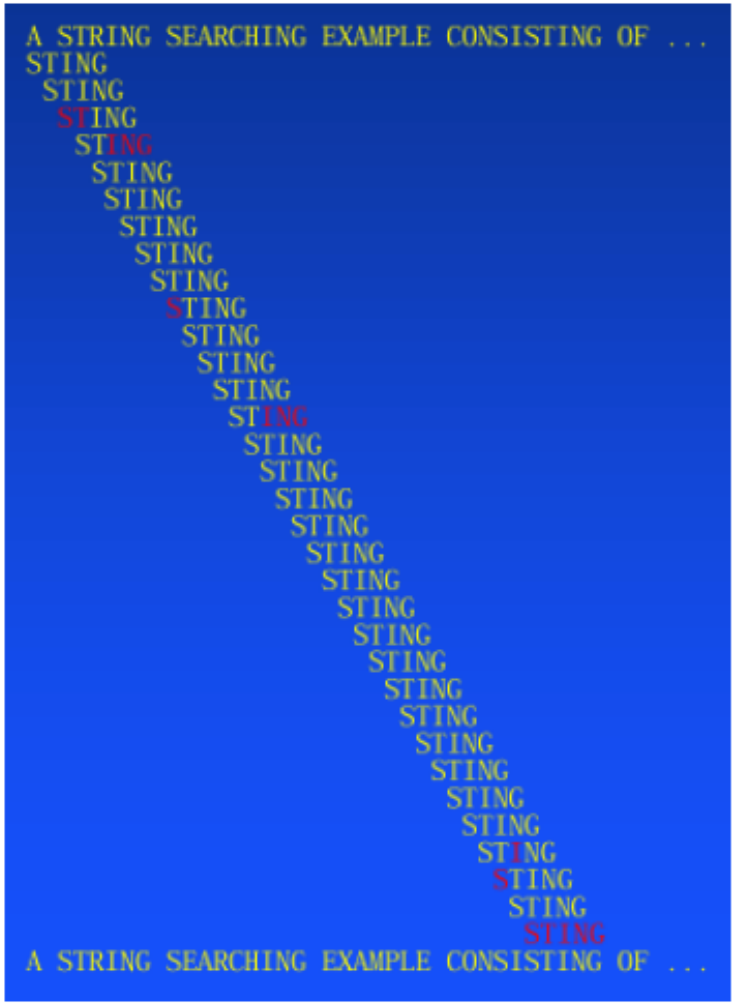


- Motif Discovery Algoritms

Motif Discovery Algorithms

Brute-Force String Matching

Brute-force String Search



Gibbs Sampling

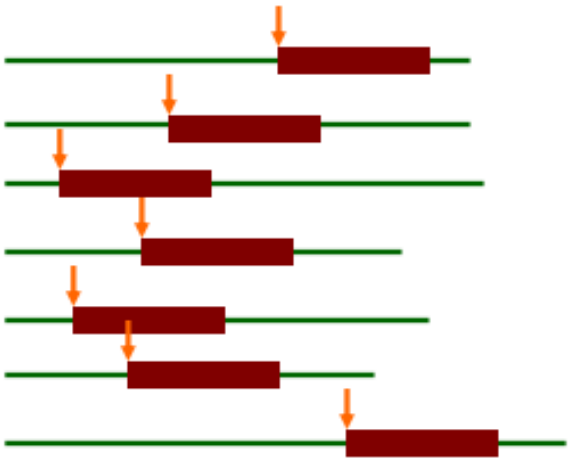
Use a simple leave-one-out sampling strategy

- First, each sequences has a motif chosen at random from it as an initial "guess".
- Next, a scoring matrix is made based on those randomly chosen sequence segments.
- Then, one sequence is chosen at random,
- And each of w-mers are compared against the initialized scoring matrix.
- The highest scoring w-mer is then throwing back into the pile of "motifs"
- And a new scoring matrix is calculated based on the new set of motifs (new because one is different).
- Repeat until "converged"
    - Then, another sequence is taken out and scanned for its highest-scoring w-mer, that w-mer is taken as that sequence's new representative motif, and it is thrown back in and a new scoring matrix is calculated.

EM Approach

The EM Approach

- 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



The Meme Algorithm

The MEME Algorithm

- Bailey & Elkan, 1993
- uses EM algorithm to find multiple motifs in a set of sequences
- first EM approach to motif discovery: Lawrence & Reilly 1990

Representing Motifs

- a motif is assumed to have a fixed width, W
- a motif is represented by a matrix of probabilities: p_{ck} represents the probability of character c in column k
- example: DNA motif with W=3

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

Background Model

Representing Motifs

- we will also represent the "background" (i.e. outside the motif) probability of each character
- p_{c0} represents the probability of character c in the background
- example:

$$p_0 = \begin{matrix} \texttt{A} & \texttt{0.26} \\ \texttt{C} & \texttt{0.24} \\ \texttt{G} & \texttt{0.23} \\ \texttt{T} & \texttt{0.27} \end{matrix}$$

Basic EM Approach

Basic EM Approach

- the element Z_{ij} of the matrix Z represents the probability that the motif starts in position j in sequence I
- example: given 4 DNA sequences of length 6, where W=3

$$Z = \begin{array}{c} \\ \text{seq1} \\ \text{seq2} \\ \text{seq3} \\ \text{seq4} \end{array} \begin{array}{cccc} \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\ 0.1 & 0.1 & 0.2 & 0.6 \\ 0.4 & 0.2 & 0.1 & 0.3 \\ 0.3 & 0.1 & 0.5 & 0.1 \\ 0.1 & 0.5 & 0.1 & 0.3 \end{array}$$

Basic EM Approach Algorithm

Basic EM Approach

given: length parameter W, training set of sequences

set initial values for p

do

re-estimate Z from p (E –step)

re-estimate p from Z (M-step)

until change in $p < \varepsilon$

return: p, Z