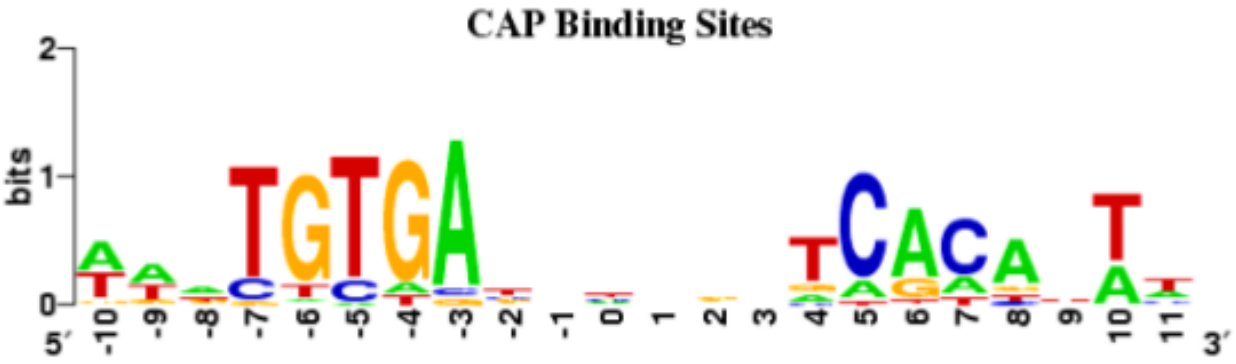
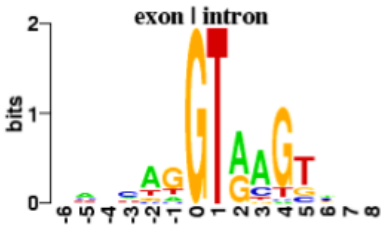
Sequence logos

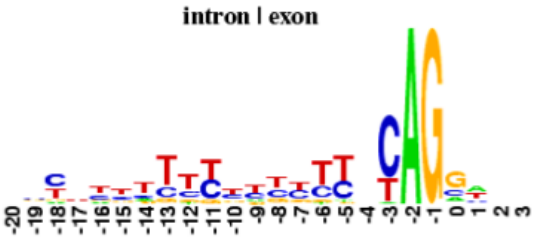
- Visual representation of positional matrices and “simple” HMM profiles
- Height of each character is proportional to its information content

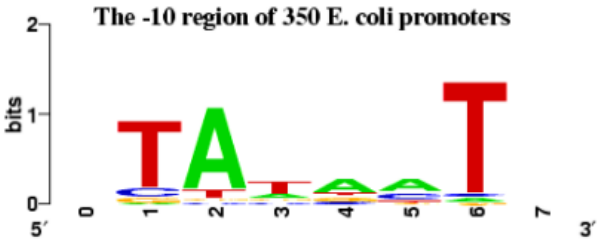


Sequence logos

- 2 bits if 1 base occurs in all input sequences
- 1 bit if two bases occur 50%
- 0 bits if all bases occur equally







Sequence logos

- Height of base b at position l

$$f(b,l)R_{sequence}(l)$$

- where

$$\frac{1}{\ln 2}\times\frac{4-1}{2n}$$

$$R_{sequence}(l) = 2 - (H(l) + e(n))$$

$$H(l) = -\sum_{b=a}^{t} f(b,l)\log_2 f(b,l)$$

Shannon entropy

Schneider and Stephens, NAR, 1990

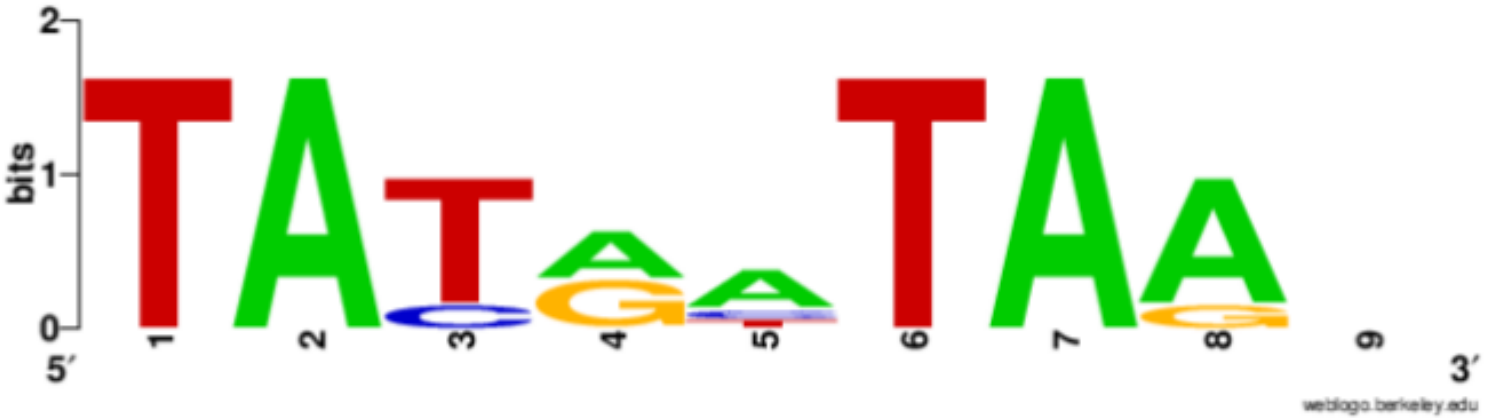
Motifs identification

- Two scenarios
  - Find known motifs (**pattern matching**)
    - Protein X binds the region upstream gene Y
    - The binding is significant?
  - Discover new motifs (**pattern discovery**)
    - Which are the motifs upstream of gene Y?
    - Which is the structure of these motifs?

Motifs identification

- Known motifs are stored in online databases
  - Multicellular organisms:
    - Transfac
    - Pazar
    - Jaspar
  - Yeast:
    - Yeastract
    - SCPD
  - Procariots:
    - RegulonDB
    - Prodoric
  - Other
    - UniProbe

Pattern matching



TATCAACATGTCGTATACCAACCTTCAACCATGTCTCAACATGTCGCG
GGTGTGCCTCCGGACCATGTCTAAGGGGTGTAAGGGTACTAACGAA
TCGTAGCATGTCCAGAGGTGCGGAGTACGTAAGGAGGGTGCCCAT
ACATGTCCGTTTCATATGAGTGCGCCTGCATTAATGTACCAACCTTCA
ACCATGTCTCAACATGTCGCGGGTGTGCCTCCACGTACGAGCCGG
AAGTCGACTCGCATGTCTGTAGGTGCGGAGTACGTAAGGAGGGTG
CCCATACATGTCCGTTTCATATGAGCCTG

Pattern matching

	1	2	3	4	5	6	7	8	9
A	-inf	1.38	-inf	0.69	0.99	-inf	1.38	1.20	-0.39
C	-inf	-inf	-0.39	-inf	-0.39	-inf	-inf	-inf	0.31
G	-inf	-inf	-inf	0.69	-inf	-inf	-inf	-0.39	0.31
T	1.38	-inf	1.20	-inf	-0.39	1.38	-inf	-inf	-0.39

TATCAACATGTCGTATACCAACCTTTATGATAAGCTCAACATGT

Pattern matching

PSEUDOCOUNTS!!!

	1	2	3	4	5	6	7	8	9
A	-3	1.38	-3	0.69	0.99	-3	1.38	1.20	-0.39
C	-3	-3	-0.39	-3	-0.39	-3	-3	-3	0.31
G	-3	-3	-3	0.69	-3	-3	-3	-0.39	0.31
T	1.38	-3	1.20	-3	-0.39	1.38	-3	-3	-0.39

TATCAACATGTCGTATACCAACCTTTATGATAAGCTCAACATGT

Pattern matching

	1	2	3	4	5	6	7	8	9
A	-3	1.38	-3	0.69	0.99	-3	1.38	1.20	-0.39
C	-3	-3	-0.39	-3	-0.39	-3	-3	-3	0.31
G	-3	-3	-3	0.69	-3	-3	-3	-0.39	0.31
T	1.38	-3	1.20	-3	-0.39	1.38	-3	-3	-0.39

TATCAACATGTCGTATACCAACCTTTATGATAAGCTCAACATGT

Pattern matching

	1	2	3	4	5	6	7	8	9
A	-3	**1.38**	-3	0.69	**0.99**	**-3**	1.38	**1.20**	-0.39
C	-3	-3	-0.39	**-3**	-0.39	-3	**-3**	-3	0.31
G	-3	-3	-3	0.69	-3	-3	-3	-0.39	0.31
T	**1.38**	-3	**1.20**	-3	-0.39	1.38	-3	-3	**-0.39**

TATCAACATGTCGTATACCAACCTTTATGATAAGCTCAACATGT

P (S | M) = -3.24

Pattern matching

	1	2	3	4	5	6	7	8	9
A	**-3**	1.38	-3	**0.69**	**0.99**	-3	**1.38**	1.20	-0.39
C	-3	-3	**-0.39**	-3	-0.39	**-3**	-3	-3	0.31
G	-3	-3	-3	0.69	-3	-3	-3	-0.39	**0.31**
T	1.38	**-3**	1.20	-3	-0.39	1.38	-3	**-3**	-0.39

TATCAACATGTCGTATACCAACCTTTATGATAAGCTCAACATGT

P (S | M) = -9.62

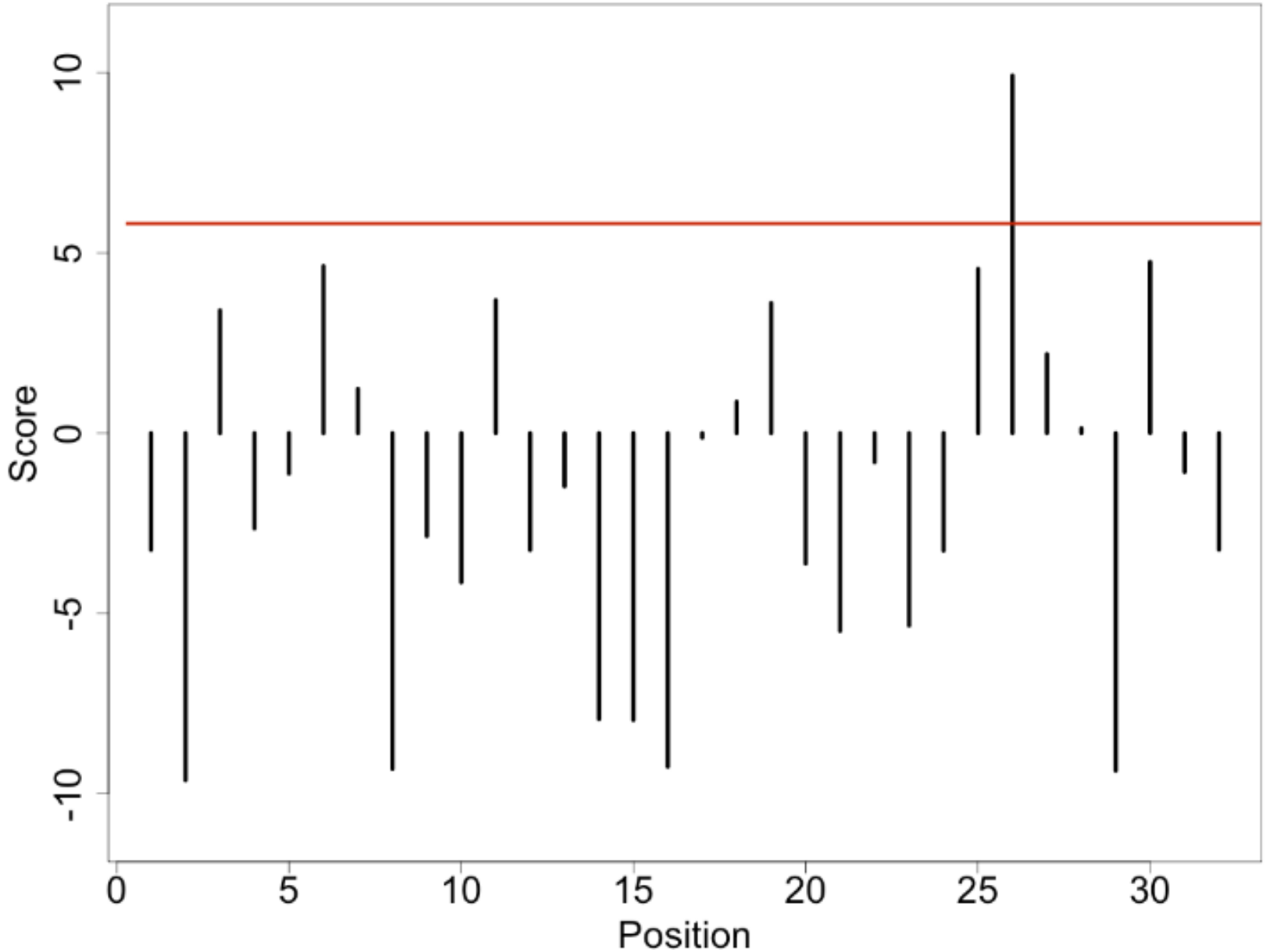
Pattern matching

	1	2	3	4	5	6	7	8	9
A	-3	**1.38**	-3	0.69	**0.99**	-3	**1.38**	**1.20**	-0.39
C	-3	-3	-0.39	-3	-0.39	-3	-3	-3	0.31
G	-3	-3	-3	**0.69**	-3	-3	-3	-0.39	**0.31**
T	**1.38**	-3	**1.20**	-3	-0.39	**1.38**	-3	-3	-0.39

TATCAACATGTCGTATACCAACCTTTATGATAAGCTCAACATGT

P (S | M) = 9.91

Pattern matching

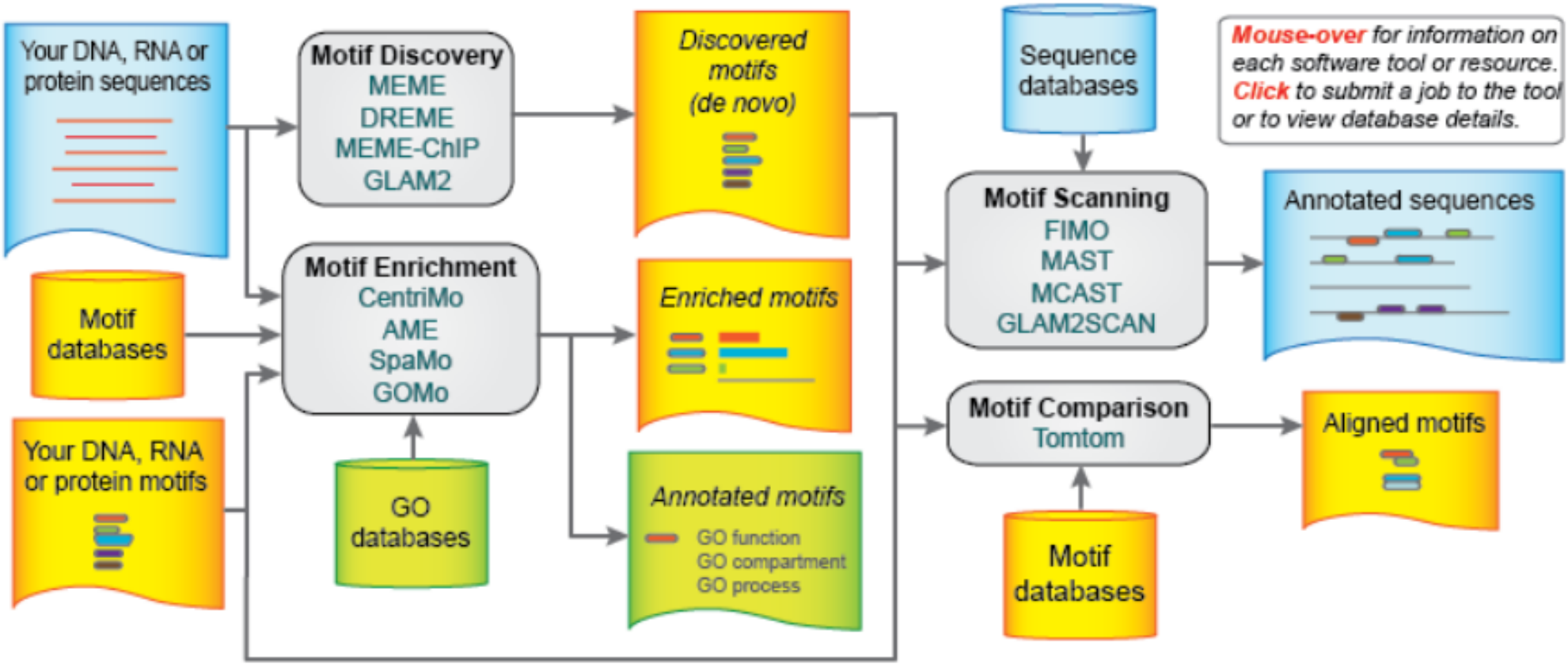


De-novo motifs identification

- Given a set of sequences
- Find the most represented motifs
- Methods:
  - Oligo-Analysis, Weeder
  - MEME
  - Gibbs sampler, MotifSampler

De-novo motifs identification

MEME suite

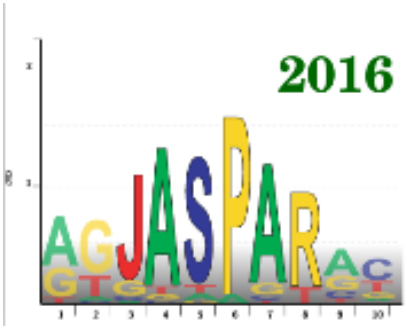


Motifs identification

Align to a custom matrix or IUPAC string ?

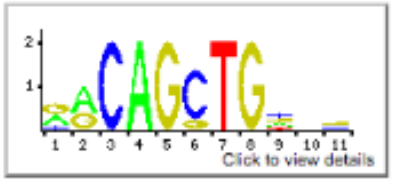
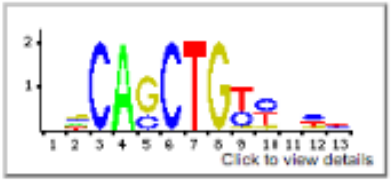
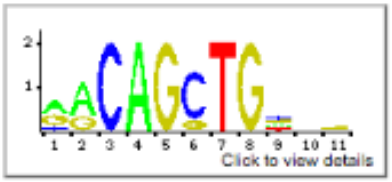
```
A [13 13  3  1 54  1  1  1  0  3  2  5 ]
C [13 39  5 53  0  1 50  1  0 37  0 17 ]
G [17  2 37  0  0 52  3  0 53  8 37 12 ]
T [11  0  9  0  0  0  0 52  1  6 15 20 ]
```

Reset | Fill in an example matrix | Align



Motifs identification

JASPAR matrix models:

TOGGLE	ID	name	species	class	family	score	percent_score	Sequence logo
☐	MA0500.1	Myog	10090	Basic helix-loop-helix factors (bHLH)	MyoD / ASC-related factors	20.65	93.8636363636364	
☐	MA0499.1	Myod1	10090	Basic helix-loop-helix factors (bHLH)	MyoD / ASC-related factors	20.5233	85.51375	
☐	MA0521.1	Tcf12	10090	Basic helix-loop-helix factors (bHLH)	E2A-related factors	20.4411	92.9140909090909	
☐	MA0697.1	ZIC3	9606	C2H2 zinc finger factors	More than 3 adjacent zinc finger factors	20.046	83.525