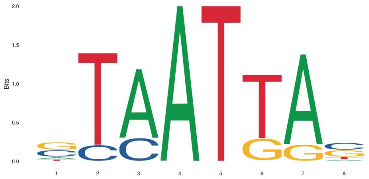
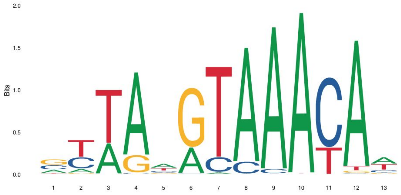
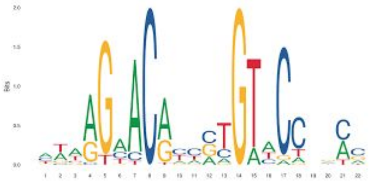
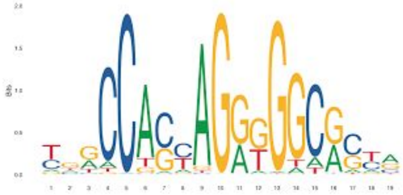
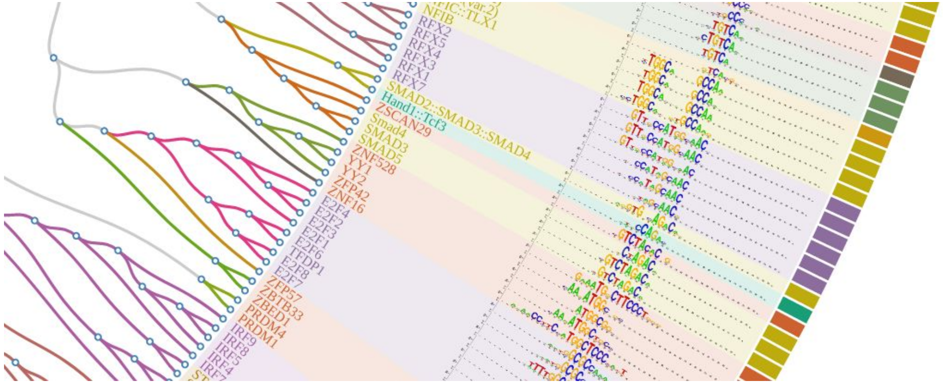








# Sequence motifs



NCMM
Centre for Molecular Medicine Norway
Nordic EMBL Partnership for Molecular Medicine

UNIVERSITAS OSLOENSIS MDCCCXI

@jaimicore
j.a.c.mondragon@ncmm.uio.no

Jaime A Castro-Mondragon
Center of Molecular Medicine Norway (NCMM)

# What is a motif ?

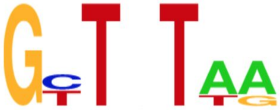
```
GCTCTTG
GCTTTAA
GTTATAA
GTTGTAA
```

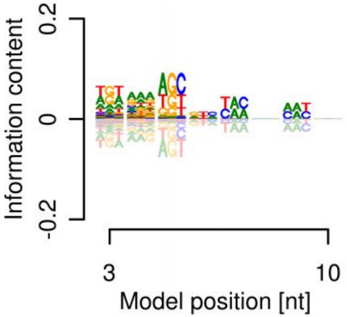
In genomics, is a pattern found in a set of biological sequences.

G[CT]T[ACTG]T[AT][AG]

- String matching
- Regular expressions
- Consensus sequences
- Position frequency matrices
- Logos
- Complex representations:
  - HMM, deep learning, matrix factorization

**GYTNTWR**





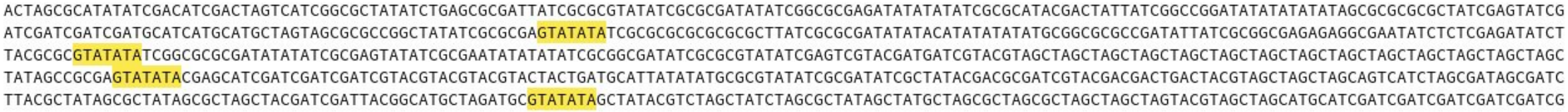
# String matching

- The simplest way to search for a string in a text.
- In DNA, we have to search in both strands, or the RC of the string pattern.

✓ Fastest option to search simple (exact) patterns

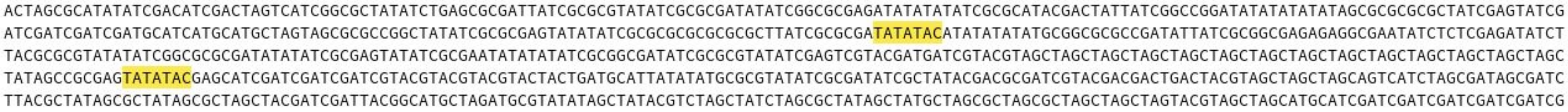
✕ Does not consider background neither nucleotide frequencies in the string.

Pattern: GTATATA



```
ACTAGCGCATATATCGACATCGACTAGTCATCGGCGCTATATCTGAGCGCGATTATCGCGCGTATATCGCGCGATATATCGGCGCGAGATATATATATCGCGCATACGACTATTATCGGCCGGATATATATATATAGCGCGCGCGCTATCGAGTATCG
ATCGATCGATCGATGCATCATGCATGCTAGTAGCGCGCCGGCTATATCGCGCGAGTATATATCGCGCGCGCGCGCTTATCGCGCGATATATACATATATATATGCGGCGCGCCGATATTATCGCGGCGAGAGAGGCGAATATCTCTCGAGATATCT
TACGCGCGTATATATCGGCGCGCGATATATATCGCGAGTATATCGCGAATATATATATCGCGGCGATATCGCGCGTATATCGAGTCGTACGATGATCGTACGTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGC
TATAGCCGCGAGTATATACGAGCATCGATCGATCGATCGTACGTACGTACGTACTACTGATGCATTATATATGCGCGTATATCGCGATATCGCTATACGACGCGATCGTACGACGACTGACTACGTAGCTAGCTAGCAGTCATCTAGCGATAGCGATC
TTACGCTATAGCGCTATAGCGCTAGCTACGATCGATTACGGCATGCTAGATGCGTATATAGCTATACGTCTAGCTATCTAGCGCTATAGCTATGCTAGCGCTAGCGCTAGCTAGCTAGTACGTAGCTAGCATGCATCGATCGATCGATCGATCGATCG
```

Pattern: TATATAC (Reverse complement)

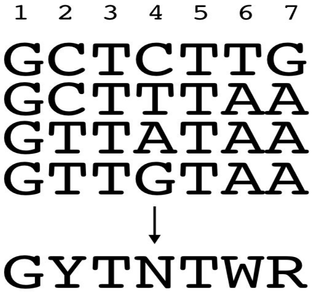


```
ACTAGCGCATATATCGACATCGACTAGTCATCGGCGCTATATCTGAGCGCGATTATCGCGCGTATATCGCGCGATATATCGGCGCGAGATATATATATCGCGCATACGACTATTATCGGCCGGATATATATATATAGCGCGCGCGCTATCGAGTATCG
ATCGATCGATCGATGCATCATGCATGCTAGTAGCGCGCCGGCTATATCGCGCGAGTATATATCGCGCGCGCGCGCTTATCGCGCGATATATACATATATATATGCGGCGCGCCGATATTATCGCGGCGAGAGAGGCGAATATCTCTCGAGATATCT
TACGCGCGTATATATCGGCGCGCGATATATATCGCGAGTATATCGCGAATATATATATCGCGGCGATATCGCGCGTATATCGAGTCGTACGATGATCGTACGTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGC
TATAGCCGCGAGTATATACGAGCATCGATCGATCGATCGTACGTACGTACGTACTACTGATGCATTATATATGCGCGTATATCGCGATATCGCTATACGACGCGATCGTACGACGACTGACTACGTAGCTAGCTAGCAGTCATCTAGCGATAGCGATC
TTACGCTATAGCGCTATAGCGCTAGCTACGATCGATTACGGCATGCTAGATGCGTATATAGCTATACGTCTAGCTATCTAGCGCTATAGCTATGCTAGCGCTAGCGCTAGCTAGCTAGTACGTAGCTAGCATGCATCGATCGATCGATCGATCGATCG
```

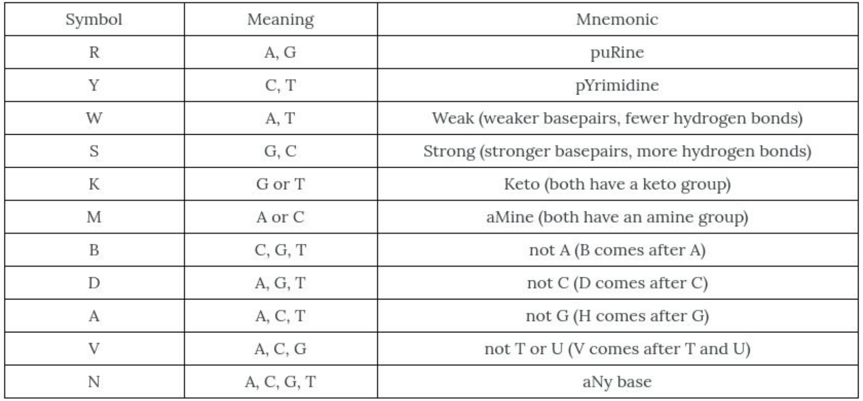
# Consensus IUPAC - Part 1

- A simple yet informative motif representation considering ambiguity.
- Regular expression for biological sequences.

✓ Represents variability/ambiguity in the string<br>
✓ Multiple sequences can be represented in a single expression<br>
× Does not consider background probabilities



```
1 2 3 4 5 6 7
GCTCTTG
GCTTTAA
GTTATAA
GTTGTAA
   ↓
GYTNTWR
```



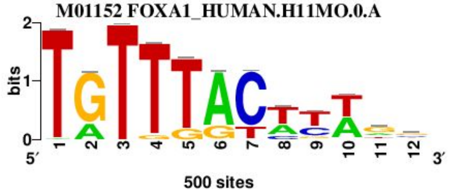
| Symbol | Meaning | Mnemonic |
|---|---|---|
| R | A, G | puRine |
| Y | C, T | pYrimidine |
| W | A, T | Weak (weaker basepairs, fewer hydrogen bonds) |
| S | G, C | Strong (stronger basepairs, more hydrogen bonds) |
| K | G or T | Keto (both have a keto group) |
| M | A or C | aMine (both have an amine group) |
| B | C, G, T | not A (B comes after A) |
| D | A, G, T | not C (D comes after C) |
| A | A, C, T | not G (H comes after G) |
| V | A, C, G | not T or U (V comes after T and U) |
| N | A, C, G, T | aNy base |

Table 2.1: IUPAC codes for nucleotides. In this table, everywhere that T applies, U applies as well.

# Consensus IUPAC - Part 2

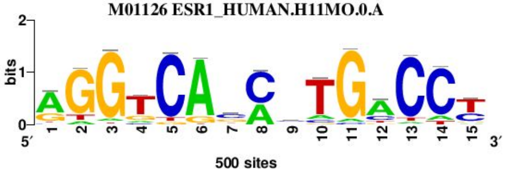
- FOXA1

| | |
|---|---|
| ○ TGTTTACWYWGS | TGTTTAC[AT][CT][AT]G[CG] |
| ○ SCWRWGTAAACA | [CG]C[AT][AG][AT]GTAAACA |



- ESR1

| | |
|---|---|
| ○ RGGTCASMSTGACCY | [AG]GGTCA[CG][AC][CG]TGACC[CT] |
| ○ RGGTCASKSTGACCY | [AG]GGTCA[CG][GT][CG]TGACC[CT] |



# Sequence complexity

- Low complexity sequences
  - Short repetitive elements
  - Many of them with biological function
  - Sequencing artifacts (contamination)

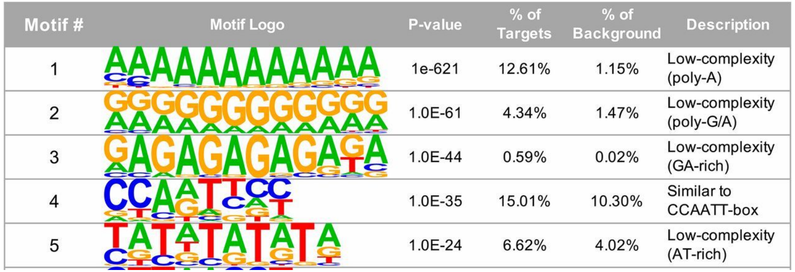
- Complexity is related to the possible number of sequences of a given string

- Confound sequence analysis

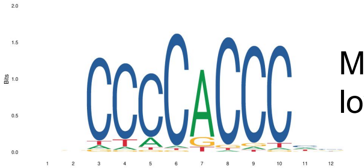
- Problematic for genome assembly

$$C_{WF} = \frac{1}{N}\log_D\left(\frac{N!}{n_A!n_C!n_G!n_T!}\right)$$

Wootton-Federhen complexity score



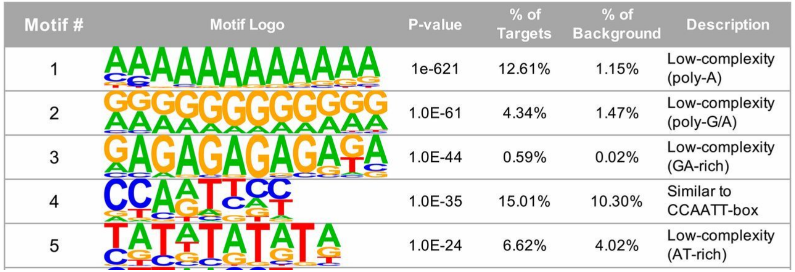
| Motif # | Motif Logo | P-value | % of Targets | % of Background | Description |
|---|---|---|---|---|---|
| 1 | AAAAAAAAAAAA | 1e-621 | 12.61% | 1.15% | Low-complexity (poly-A) |
| 2 | GGGGGGGGGGGG<br>AAAAAAAAAAAA | 1.0E-61 | 4.34% | 1.47% | Low-complexity (poly-G/A) |
| 3 | GAGAGAGAGAGA | 1.0E-44 | 0.59% | 0.02% | Low-complexity (GA-rich) |
| 4 | CCAATTCC | 1.0E-35 | 15.01% | 10.30% | Similar to CCAATT-box |
| 5 | TATATATATA | 1.0E-24 | 6.62% | 4.02% | Low-complexity (AT-rich) |



Many TFs (Kruppel-like) bind low-complexity sequences.

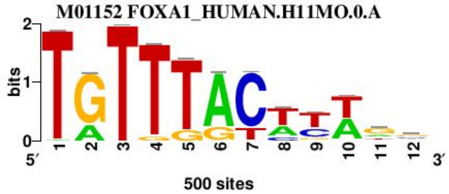
# Sequence complexity

| Name | Sequence | WF |
|---|---|---|
| Poly-A | AAAAAAAA | 0.0000000 |
| KLF4 | CCCCACCC | 0.1875000 |
| GA-rich | GAGAGAGAGA | 0.3988640 |
| FOXA1 | TGTTTACTTT | 0.4745927 |
| ESR1 | AGGTCACCCTGACCT | 0.7667800 |
| CTCF | TGGCCACCAGGGGGCGCTA | 0.7468272 |

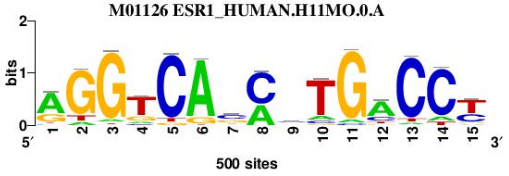


| Motif # | Motif Logo | P-value | % of Targets | % of Background | Description |
|---|---|---|---|---|---|
| 1 | | 1e-621 | 12.61% | 1.15% | Low-complexity (poly-A) |
| 2 | | 1.0E-61 | 4.34% | 1.47% | Low-complexity (poly-G/A) |
| 3 | | 1.0E-44 | 0.59% | 0.02% | Low-complexity (GA-rich) |
| 4 | | 1.0E-35 | 15.01% | 10.30% | Similar to CCAATT-box |
| 5 | | 1.0E-24 | 6.62% | 4.02% | Low-complexity (AT-rich) |

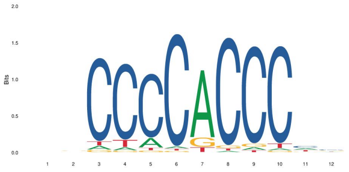
$$C_{WF} = \frac{1}{N}\log_D\left(\frac{N!}{n_A!n_C!n_G!n_T!}\right)$$

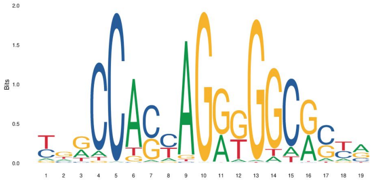


M01152 FOXA1_HUMAN.H11MO.0.A



M01126 ESR1_HUMAN.H11MO.0.A





# Position Frequency/Weight Matrices

- The most used model for biological sequences.
- Probabilistic representation of sequences.
- A simple matrix representing the nucleotide/aminoacid frequencies along a sequence.
- Represent TF binding motifs, TSSs, Core-promoter elements, Splice sites, Amino-acid domains, etc.

✓ Intuitive and simple representation

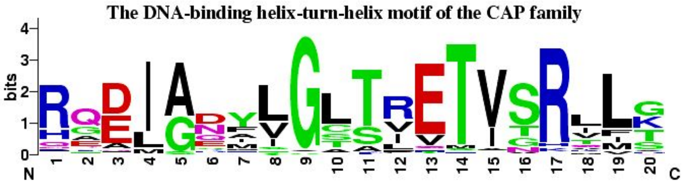
✓ Allow to integrate background frequencies

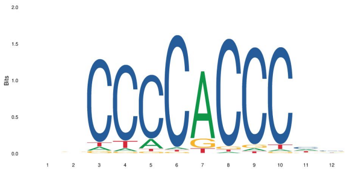
✕ Assumes independency among nucleotides/aminoacids.

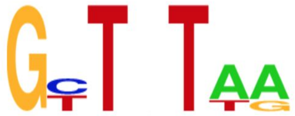
GCTCTTG
GCTTTAA
GTTATAA
GTTGTAA

↓

GYTNTWR



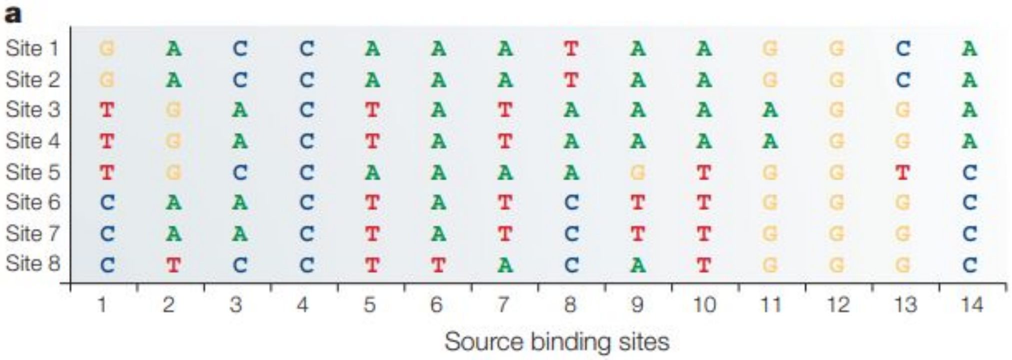




# Position Frequency/Weight Matrices

A collection of known sites, aligned and with the same length.

Experimentally validated or predicted.

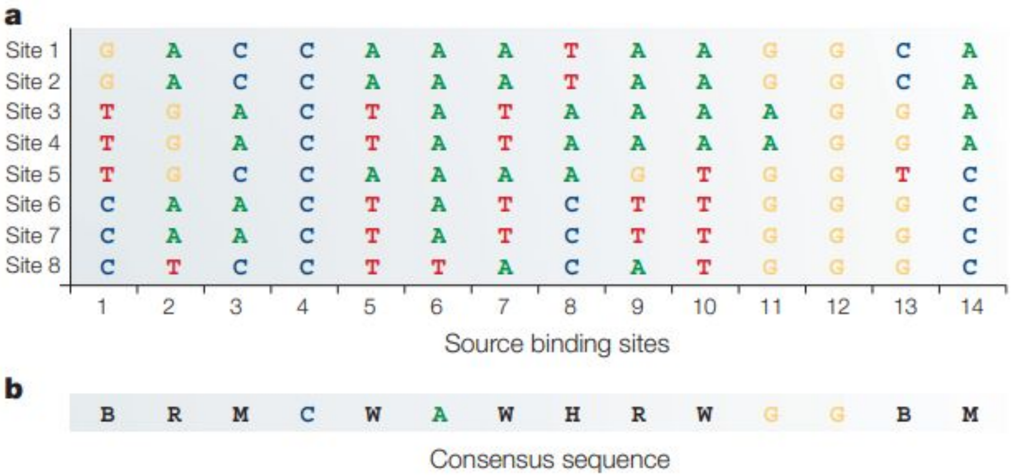


| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Site 1 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 2 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 3 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 4 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 5 | T | G | C | C | A | A | A | A | G | T | G | G | T | C |
| Site 6 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 7 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 8 | C | T | C | C | T | T | A | C | A | T | G | G | G | C |

Source binding sites

Sandelin and Wasserman. 2004

# Position Frequency/Weight Matrices



a

|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Site 1 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 2 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 3 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 4 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 5 | T | G | C | C | A | A | A | A | G | T | G | G | T | C |
| Site 6 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 7 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 8 | C | T | C | C | T | T | A | C | A | T | G | G | G | C |

Source binding sites

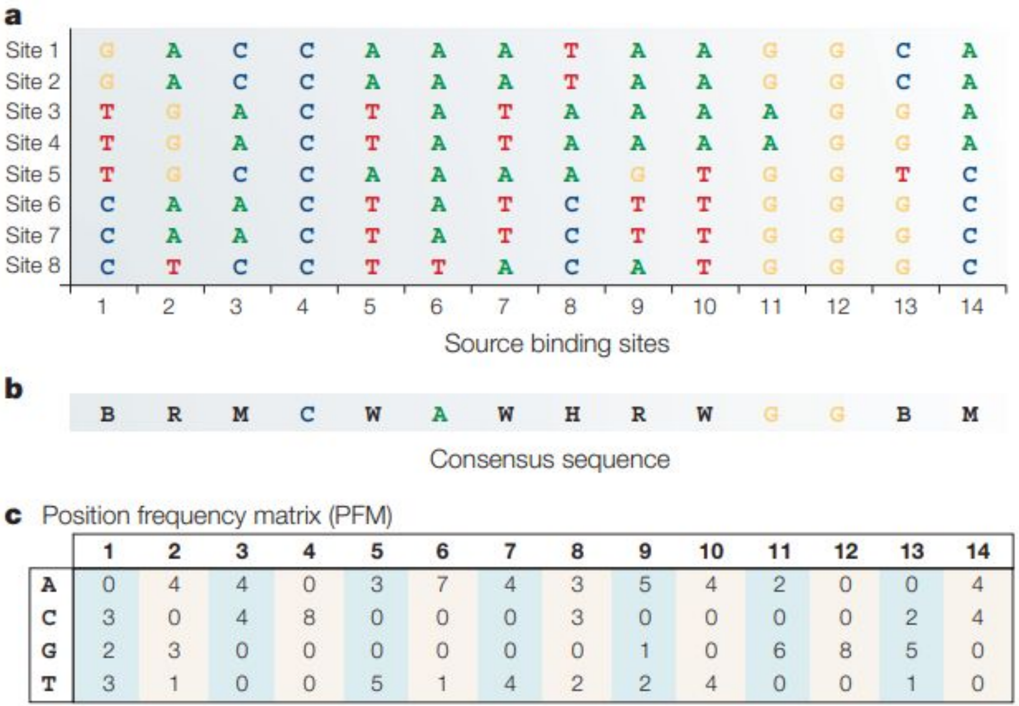
b

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| B | R | M | C | W | A | W | H | R | W | G | G | B | M |

Consensus sequence

Sandelin and Wasserman. 2004

# Position Frequency/Weight Matrices



**a**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Site 1 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 2 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 3 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 4 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 5 | T | G | C | C | A | A | A | A | G | T | G | G | T | C |
| Site 6 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 7 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 8 | C | T | C | C | T | T | A | C | A | T | G | G | G | C |

Source binding sites

**b**

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| B | R | M | C | W | A | W | H | R | W | G | G | B | M |

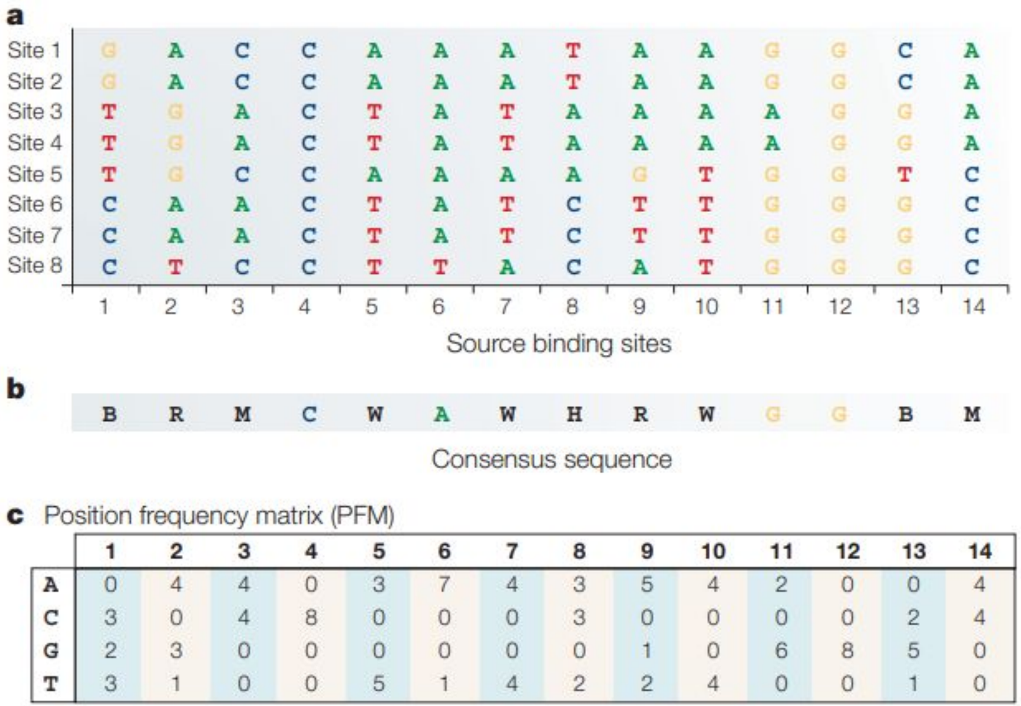
Consensus sequence

**c** Position frequency matrix (PFM)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 4 | 4 | 0 | 3 | 7 | 4 | 3 | 5 | 4 | 2 | 0 | 0 | 4 |
| C | 3 | 0 | 4 | 8 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 2 | 4 |
| G | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 6 | 8 | 5 | 0 |
| T | 3 | 1 | 0 | 0 | 5 | 1 | 4 | 2 | 2 | 4 | 0 | 0 | 1 | 0 |

Sandelin and Wasserman. 2004

# Position Frequency/Weight Matrices



**a**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Site 1 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 2 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 3 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 4 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 5 | T | G | C | C | A | A | A | A | G | T | G | G | T | C |
| Site 6 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 7 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 8 | C | T | C | C | T | T | A | C | A | T | G | G | G | C |

Source binding sites

**b**

B R M C W A W H R W G G B M

Consensus sequence

**c** Position frequency matrix (PFM)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 4 | 4 | 0 | 3 | 7 | 4 | 3 | 5 | 4 | 2 | 0 | 0 | 4 |
| C | 3 | 0 | 4 | 8 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 2 | 4 |
| G | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 6 | 8 | 5 | 0 |
| T | 3 | 1 | 0 | 0 | 5 | 1 | 4 | 2 | 2 | 4 | 0 | 0 | 1 | 0 |

Sandelin and Wasserman. 2004

PFM databases (for TF binding motifs)

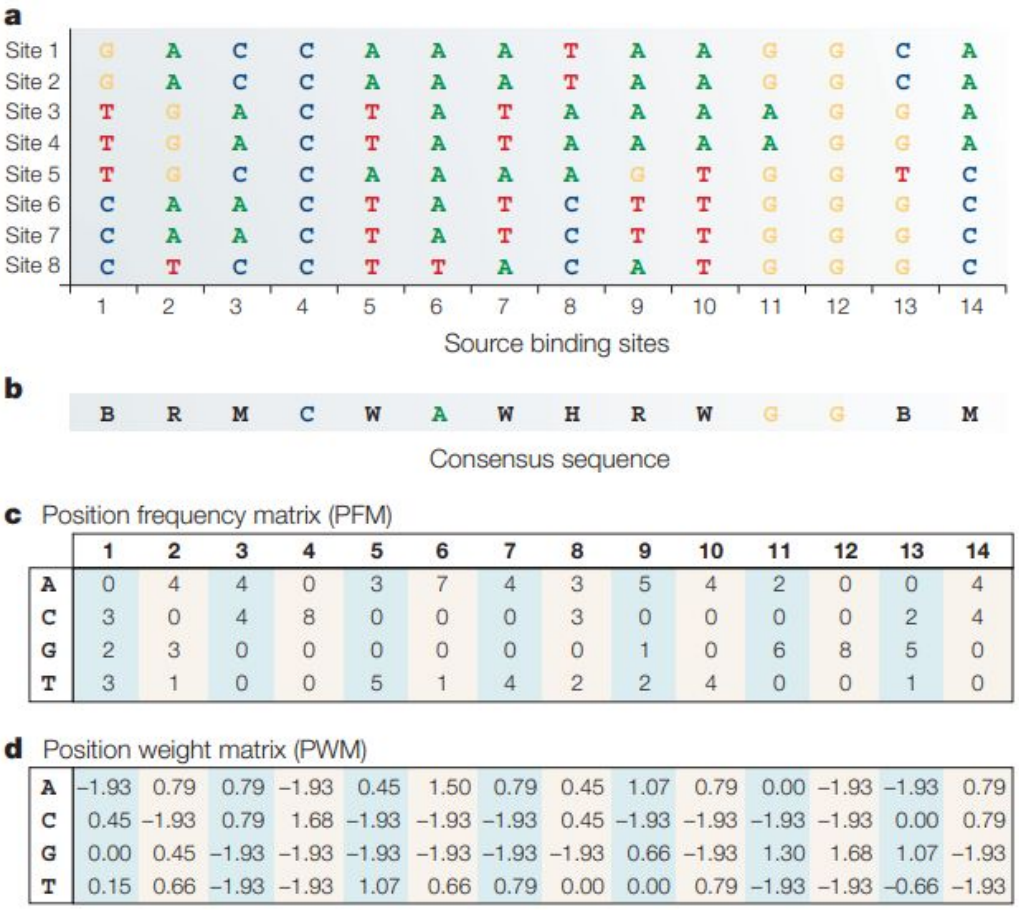
JASPAR 2018

HOCOMOCO

CIS-BP

RegulonDB

# Position Frequency/Weight Matrices



**a**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Site 1 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 2 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 3 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 4 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 5 | T | G | C | C | A | A | A | A | G | T | G | G | T | C |
| Site 6 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 7 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 8 | C | T | C | C | T | T | A | C | A | T | G | G | G | C |

Source binding sites

**b**

B R M C W A W H R W G G B M

Consensus sequence

**c** Position frequency matrix (PFM)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 4 | 4 | 0 | 3 | 7 | 4 | 3 | 5 | 4 | 2 | 0 | 0 | 4 |
| C | 3 | 0 | 4 | 8 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 2 | 4 |
| G | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 6 | 8 | 5 | 0 |
| T | 3 | 1 | 0 | 0 | 5 | 1 | 4 | 2 | 2 | 4 | 0 | 0 | 1 | 0 |

**d** Position weight matrix (PWM)

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | –1.93 | 0.79 | 0.79 | –1.93 | 0.45 | 1.50 | 0.79 | 0.45 | 1.07 | 0.79 | 0.00 | –1.93 | –1.93 | 0.79 |
| C | 0.45 | –1.93 | 0.79 | 1.68 | –1.93 | –1.93 | –1.93 | 0.45 | –1.93 | –1.93 | –1.93 | –1.93 | 0.00 | 0.79 |
| G | 0.00 | 0.45 | –1.93 | –1.93 | –1.93 | –1.93 | –1.93 | –1.93 | 0.66 | –1.93 | 1.30 | 1.68 | 1.07 | –1.93 |
| T | 0.15 | 0.66 | –1.93 | –1.93 | 1.07 | 0.66 | 0.79 | 0.00 | 0.00 | 0.79 | –1.93 | –1.93 | –0.66 | –1.93 |

The transition from Frequencies to Weights requires a background.

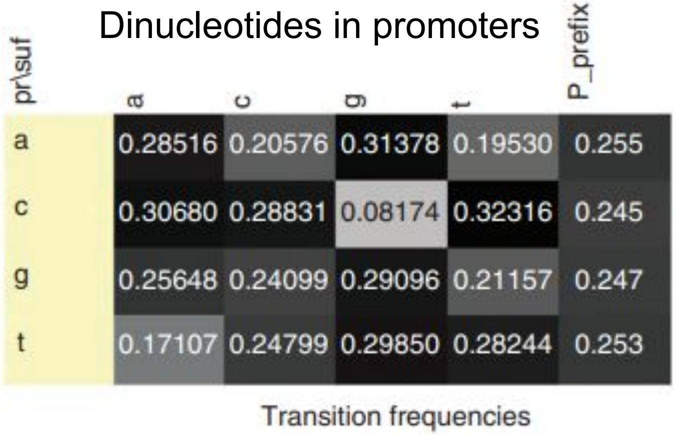
Sandelin and Wasserman. 2004

# Background Models

- The expected frequencies of nucleotides/aminoacids.
- Indicates what are the sequences with higher probability to appear by chance.
- The values vary according the sequence context: promoters, exons, CpG islands.
- Can be modeled using markov chains of higher order (di-, tri- , k-nucleotides)

- Human genome (single nucleotide):
    - A: 0.204
    - C: 0.295
    - G: 0.296
    - T : 0.205

Dinucleotides in promoters



| pr\suf | a | c | g | t | P_prefix |
|---|---|---|---|---|---|
| a | 0.28516 | 0.20576 | 0.31378 | 0.19530 | 0.255 |
| c | 0.30680 | 0.28831 | 0.08174 | 0.32316 | 0.245 |
| g | 0.25648 | 0.24099 | 0.29096 | 0.21157 | 0.247 |
| t | 0.17107 | 0.24799 | 0.29850 | 0.28244 | 0.253 |

Transition frequencies

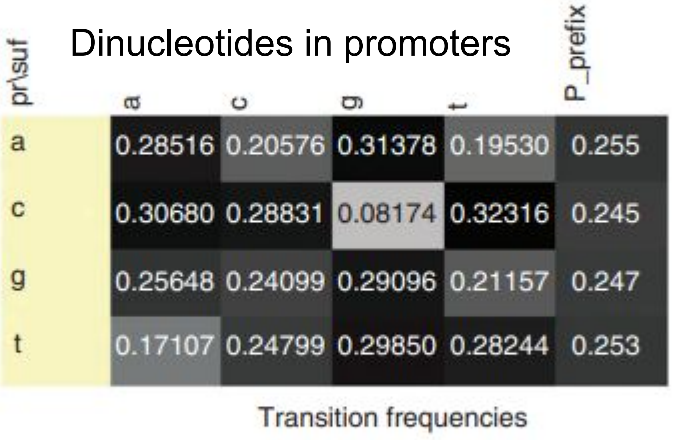
# Background Models

- Probability of ATACGT
  - Single: (0.204 ^ 2) * (0.205 ^ 2) * 0.295 * 0.296 = 1.527e-04
  - Dinucleotide : 0.204 * 0.195 * 0.171 * 0.205 * 0.081 * 0.211 = 2.383e-05
    A (AT) (TA) (AC) (CG) (GT)

Human genome (single nucleotide):

- A: 0.204
- C: 0.295
- G: 0.296
- T : 0.205

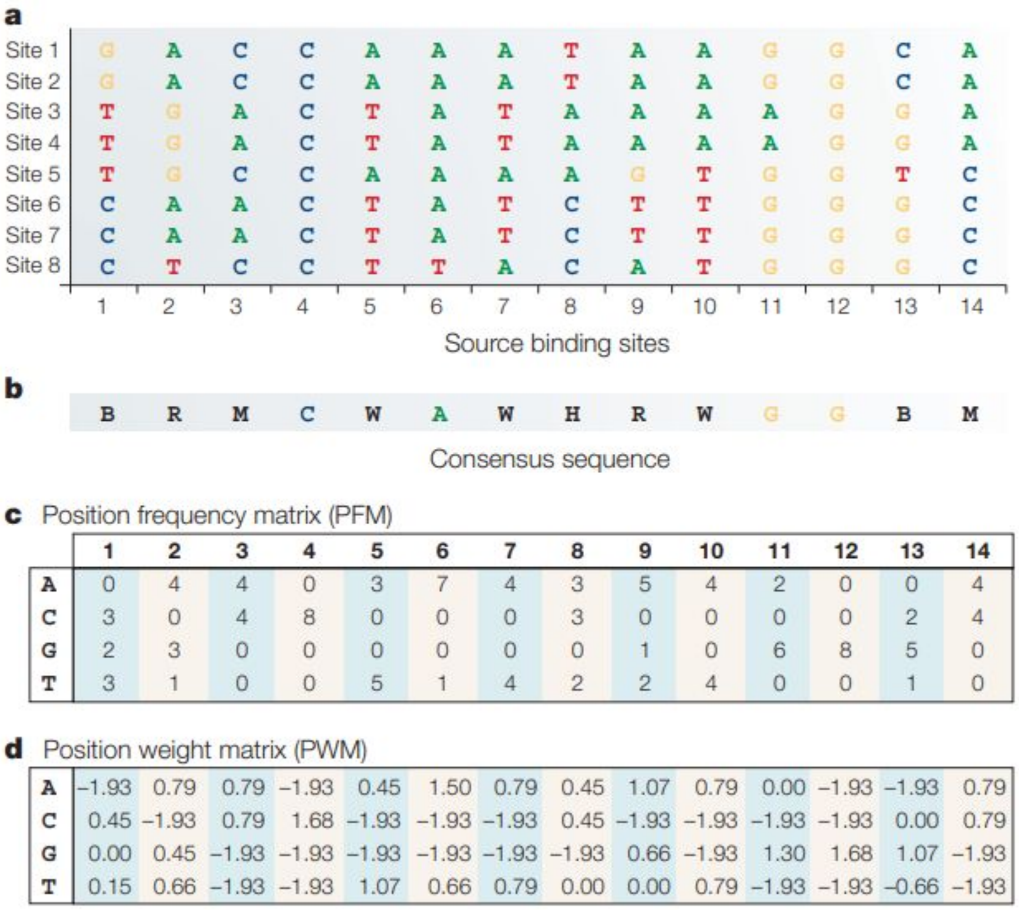
Dinucleotides in promoters



| pr\suf | a | c | g | t | P_prefix |
|---|---|---|---|---|---|
| a | 0.28516 | 0.20576 | 0.31378 | 0.19530 | 0.255 |
| c | 0.30680 | 0.28831 | 0.08174 | 0.32316 | 0.245 |
| g | 0.25648 | 0.24099 | 0.29096 | 0.21157 | 0.247 |
| t | 0.17107 | 0.24799 | 0.29850 | 0.28244 | 0.253 |

Transition frequencies

# Position Frequency/Weight Matrices



**a**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Site 1 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 2 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 3 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 4 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 5 | T | G | C | C | A | A | A | A | G | T | G | G | T | C |
| Site 6 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 7 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 8 | C | T | C | C | T | T | A | C | A | T | G | G | G | C |

Source binding sites

**b**

| B | R | M | C | W | A | W | H | R | W | G | G | B | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

Consensus sequence

**c** Position frequency matrix (PFM)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 4 | 4 | 0 | 3 | 7 | 4 | 3 | 5 | 4 | 2 | 0 | 0 | 4 |
| C | 3 | 0 | 4 | 8 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 2 | 4 |
| G | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 6 | 8 | 5 | 0 |
| T | 3 | 1 | 0 | 0 | 5 | 1 | 4 | 2 | 2 | 4 | 0 | 0 | 1 | 0 |

**d** Position weight matrix (PWM)

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | −1.93 | 0.79 | 0.79 | −1.93 | 0.45 | 1.50 | 0.79 | 0.45 | 1.07 | 0.79 | 0.00 | −1.93 | −1.93 | 0.79 |
| C | 0.45 | −1.93 | 0.79 | 1.68 | −1.93 | −1.93 | −1.93 | 0.45 | −1.93 | −1.93 | −1.93 | −1.93 | 0.00 | 0.79 |
| G | 0.00 | 0.45 | −1.93 | −1.93 | −1.93 | −1.93 | −1.93 | −1.93 | 0.66 | −1.93 | 1.30 | 1.68 | 1.07 | −1.93 |
| T | 0.15 | 0.66 | −1.93 | −1.93 | 1.07 | 0.66 | 0.79 | 0.00 | 0.00 | 0.79 | −1.93 | −1.93 | −0.66 | −1.93 |

Sandelin and Wasserman. 2004

$$\boxed{W_{i,j} = \ln\left(\frac{f'_{i,j}}{p_i}\right)}$$

f : probabilities of each nucleotide in the PFM
p : background frequencies

# Position Frequency/Weight Matrices

**a**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Site 1 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 2 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 3 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 4 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 5 | T | G | C | C | A | A | A | A | G | T | G | G | T | C |
| Site 6 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 7 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 8 | C | T | C | C | T | T | A | C | A | T | G | G | G | C |

Source binding sites

**b**

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| B | R | M | C | W | A | W | H | R | W | G | G | B | M |

Consensus sequence

**c** Position frequency matrix (PFM)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 4 | 4 | 0 | 3 | 7 | 4 | 3 | 5 | 4 | 2 | 0 | 0 | 4 |
| C | 3 | 0 | 4 | 8 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 2 | 4 |
| G | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 6 | 8 | 5 | 0 |
| T | 3 | 1 | 0 | 0 | 5 | 1 | 4 | 2 | 2 | 4 | 0 | 0 | 1 | 0 |

**d** Position weight matrix (PWM)

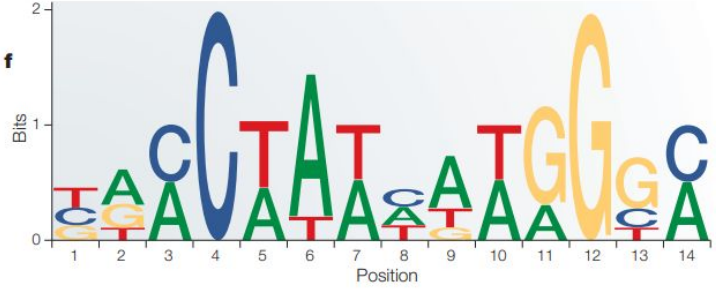
| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | –1.93 | 0.79 | 0.79 | –1.93 | 0.45 | 1.50 | 0.79 | 0.45 | 1.07 | 0.79 | 0.00 | –1.93 | –1.93 | 0.79 |
| C | 0.45 | –1.93 | 0.79 | 1.68 | –1.93 | –1.93 | –1.93 | 0.45 | –1.93 | –1.93 | –1.93 | –1.93 | 0.00 | 0.79 |
| G | 0.00 | 0.45 | –1.93 | –1.93 | –1.93 | –1.93 | –1.93 | –1.93 | 0.66 | –1.93 | 1.30 | 1.68 | 1.07 | –1.93 |
| T | 0.15 | 0.66 | –1.93 | –1.93 | 1.07 | 0.66 | 0.79 | 0.00 | 0.00 | 0.79 | –1.93 | –1.93 | –0.66 | –1.93 |

**e** Site scoring

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.45 | –0.66 | 0.79 | 1.68 | 0.45 | –0.66 | 0.79 | 0.45 | –0.66 | 0.79 | 0.00 | 1.68 | –0.66 | 0.79 |
| T | T | A | C | A | T | A | A | G | T | A | G | T | C |

Σ = 5.23, 78% of maximum

**f**



Sandelin and Wasserman. 2004

# Position Frequency/Weight Matrices

**c** Position frequency matrix (PFM)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 4 | 4 | 0 | 3 | 7 | 4 | 3 | 5 | 4 | 2 | 0 | 0 | 4 |
| C | 3 | 0 | 4 | 8 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 2 | 4 |
| G | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 6 | 8 | 5 | 0 |
| T | 3 | 1 | 0 | 0 | 5 | 1 | 4 | 2 | 2 | 4 | 0 | 0 | 1 | 0 |

**d** Position weight matrix (PWM)

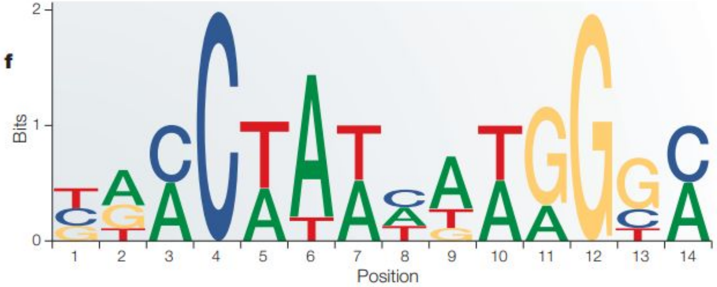
| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | –1.93 | 0.79 | 0.79 | –1.93 | 0.45 | 1.50 | 0.79 | 0.45 | 1.07 | 0.79 | 0.00 | –1.93 | –1.93 | 0.79 |
| C | 0.45 | –1.93 | 0.79 | 1.68 | –1.93 | –1.93 | –1.93 | 0.45 | –1.93 | –1.93 | –1.93 | –1.93 | 0.00 | 0.79 |
| G | 0.00 | 0.45 | –1.93 | –1.93 | –1.93 | –1.93 | –1.93 | –1.93 | 0.66 | –1.93 | 1.30 | 1.68 | 1.07 | –1.93 |
| T | 0.15 | 0.66 | –1.93 | –1.93 | 1.07 | 0.66 | 0.79 | 0.00 | 0.00 | 0.79 | –1.93 | –1.93 | –0.66 | –1.93 |

**e** Site scoring

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.45 | –0.66 | 0.79 | 1.68 | 0.45 | –0.66 | 0.79 | 0.45 | –0.66 | 0.79 | 0.00 | 1.68 | –0.66 | 0.79 |
| T | T | A | C | A | T | A | A | G | T | A | G | T | C |

$\Sigma = 5.23$, 78% of maximum

**f**



Probability given the Frequency matrix

$$S(x_j) = \log\left(\frac{\boxed{P(x|M)}}{\boxed{P(x|R)}}\right) = \sum_{i=1}^{K} \log\left(\frac{\boxed{f_{i x_j[i]}}}{\boxed{p_{x_j[i]}}}\right)$$

Probability given the Background model

Sandelin and Wasserman. 2004

# Position Frequency/Weight Matrices

**a**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Site 1 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 2 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 3 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 4 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 5 | T | G | C | C | A | A | A | A | G | T | G | G | T | C |
| Site 6 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 7 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 8 | C | T | C | C | T | T | A | C | A | T | G | G | G | C |

Source binding sites

**b**

B R M C W A W H R W G G B M

Consensus sequence

**c** Position frequency matrix (PFM)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 4 | 4 | 0 | 3 | 7 | 4 | 3 | 5 | 4 | 2 | 0 | 0 | 4 |
| C | 3 | 0 | 4 | 8 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 2 | 4 |
| G | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 6 | 8 | 5 | 0 |
| T | 3 | 1 | 0 | 0 | 5 | 1 | 4 | 2 | 2 | 4 | 0 | 0 | 1 | 0 |

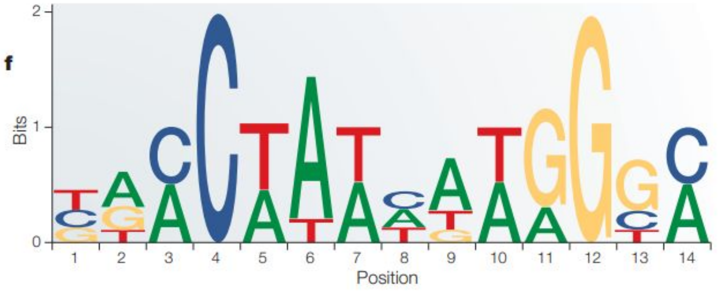
**d** Position weight matrix (PWM)

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | –1.93 | 0.79 | 0.79 | –1.93 | 0.45 | 1.50 | 0.79 | 0.45 | 1.07 | 0.79 | 0.00 | –1.93 | –1.93 | 0.79 |
| C | 0.45 | –1.93 | 0.79 | 1.68 | –1.93 | –1.93 | –1.93 | 0.45 | –1.93 | –1.93 | –1.93 | –1.93 | 0.00 | 0.79 |
| G | 0.00 | 0.45 | –1.93 | –1.93 | –1.93 | –1.93 | –1.93 | –1.93 | 0.66 | –1.93 | 1.30 | 1.68 | 1.07 | –1.93 |
| T | 0.15 | 0.66 | –1.93 | –1.93 | 1.07 | 0.66 | 0.79 | 0.00 | 0.00 | 0.79 | –1.93 | –1.93 | –0.66 | –1.93 |

**e** Site scoring

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.45 | –0.66 | 0.79 | 1.68 | 0.45 | –0.66 | 0.79 | 0.45 | –0.66 | 0.79 | 0.00 | 1.68 | –0.66 | 0.79 |
| T | T | A | C | A | T | A | A | G | T | A | G | T | C |

Σ = 5.23, 78% of maximum



Sites with W > 0 could be potential binding sites

Max score: 14.22

Sandelin and Wasserman. 2004

# Information Content (IC)

$$H_g = -\sum_{i=1}^{A} p_i \log_2(p_i)$$

- Shannon entropy is a measure of the uncertainty of a model

- Special cases of uncertainty (for a 4 letter alphabet):

  - **min(H) = 0**
    No uncertainty at all: the nucleotide is completely specified (e.g. p={1, 0, 0, 0})

  - **H=1**
    Uncertainty between two letters (e.g. p= {0.5, 0, 0, 0.5})

  - **max(H) = 2**
    Complete uncertainty (e.g. p= {0.25, 0.25, 0.25, 0.25})

$$H_{max} = -\left(\tfrac{1}{4}\log_2\left(\tfrac{1}{4}\right) + \tfrac{1}{4}\log_2\left(\tfrac{1}{4}\right) + \tfrac{1}{4}\log_2\left(\tfrac{1}{4}\right) + \tfrac{1}{4}\log_2\left(\tfrac{1}{4}\right)\right)$$

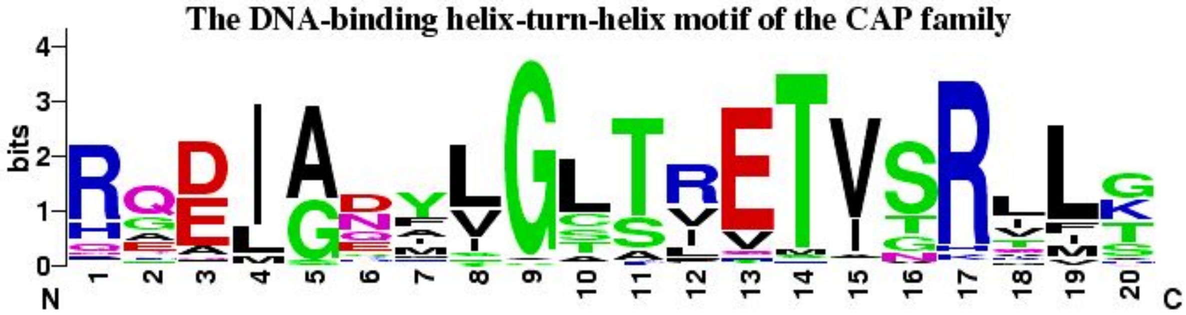
- Information content (IC) of a PFM is the sum of the differences between the Max Entropy and the observed entropy on each column.
  - IC = $H_{max}$ - $H_g$

# Information Content (IC)

$$H_g = -\sum_{i=1}^{A} p_i \log_2(p_i)$$

- Shannon entropy is a measure of the uncertainty of a model
- Special cases of uncertainty (for a 20 letter alphabet, aminoacids):
  - **max(H) = 4.32**

    $H_{max}$ = -( 1/20 * $\log_2$(1/20) ) * 20

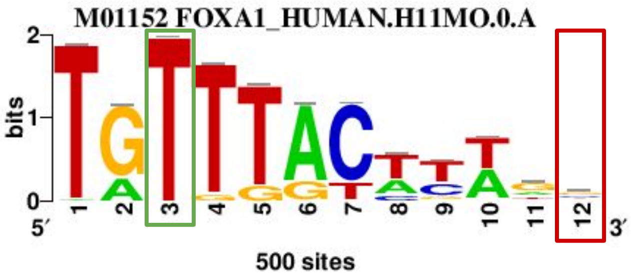


The DNA-binding helix-turn-helix motif of the CAP family

# Information Content (IC) - Examples

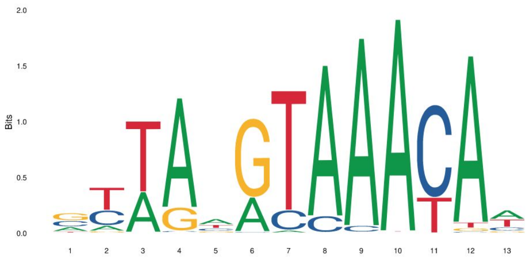
|   |     |     |     |     |     |     |     |     |     |     |     |     |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| A | 8 | 123 | 1 | 2 | 0 | 391 | 4 | 159 | 16 | 232 | 109 | 54 |
| C | 0 | 2 | 0 | 0 | 0 | 0 | 399 | 69 | 157 | 10 | 49 | 130 |
| G | 0 | 375 | 0 | 28 | 72 | 103 | 4 | 2 | 54 | 8 | 247 | 197 |
| T | 492 | 0 | 499 | 470 | 428 | 6 | 93 | 270 | 273 | 250 | 95 | 119 |
|   | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |

$$H_g = -\sum_{i=1}^{A} p_i \log_2(p_i)$$

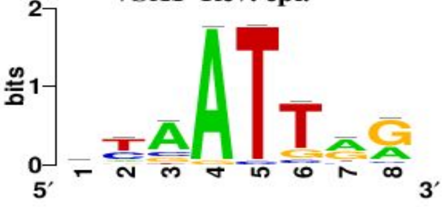


- IC = $H_{max}$ - $H_g$
- **IC$_3$**: $2 + (\ 1/500 * \log_2(1/500)\ ) + 0 + 0 + (\ 499/500 * \log_2(499/500)\ ) = \boxed{1.97}$
- **IC$_{12}$**: $2 + (\ 54/500 * \log_2(54/500)\ ) + (\ 130/500 * \log_2(130/500)\ ) + (\ 197/500 * \log_2(197/500)\ ) + (\ 119/500 * \log_2(119/500)\ ) = \boxed{0.12}$

# Information Content (IC) - Examples

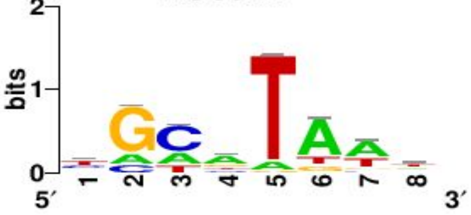


VSX1 Rev. cpl.

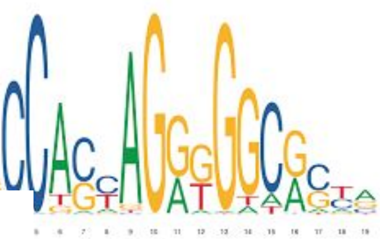


3028 sites

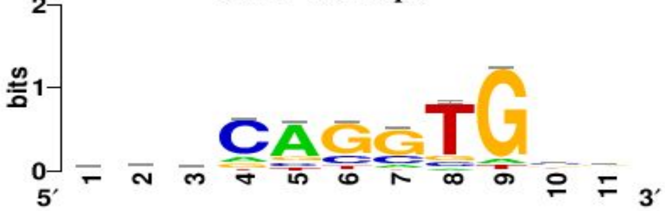
POU3F4



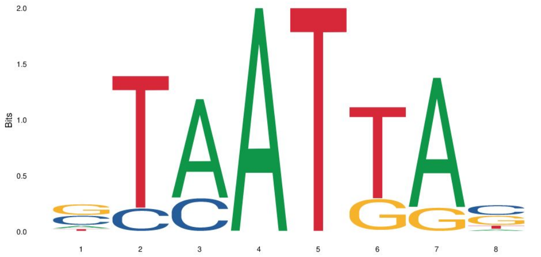
567 sites

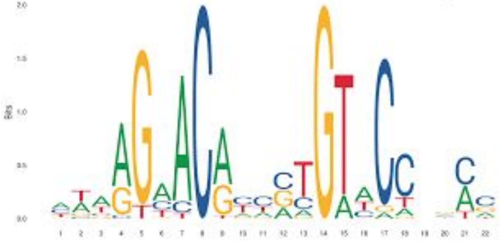


TFEB Rev. cpl.



467 sites





# Motif analysis algorithms

https://github.com/daquang/YAMDA

http://autosome.ru/ChIPMunk/

http://rsat-tagc.univ-mrs.fr/rsat/

http://homer.ucsd.edu/homer/motif/

https://meme-suite.org/meme/

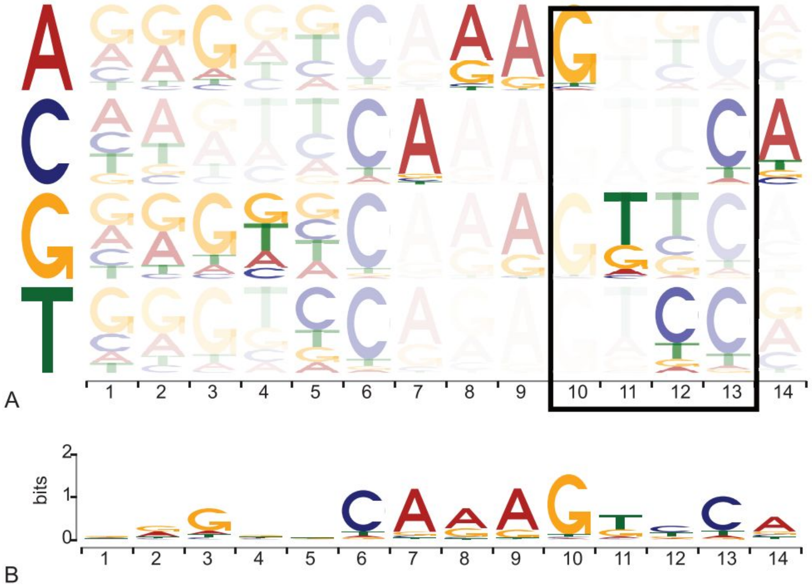
GimmeMotifs

https://github.com/vanheeringen-lab/gimmemotifs

# Complex representation of motifs

**HMM**

Mathelier 2013



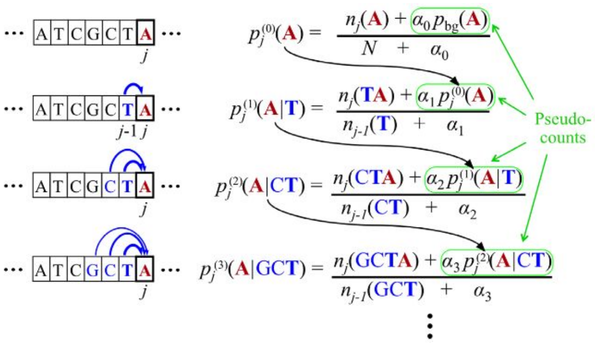
Nucleotide dependencies are modelled

# Complex representation of motifs

Supports long k-mers

**Bayesian Markov Models**

Siebert 2016



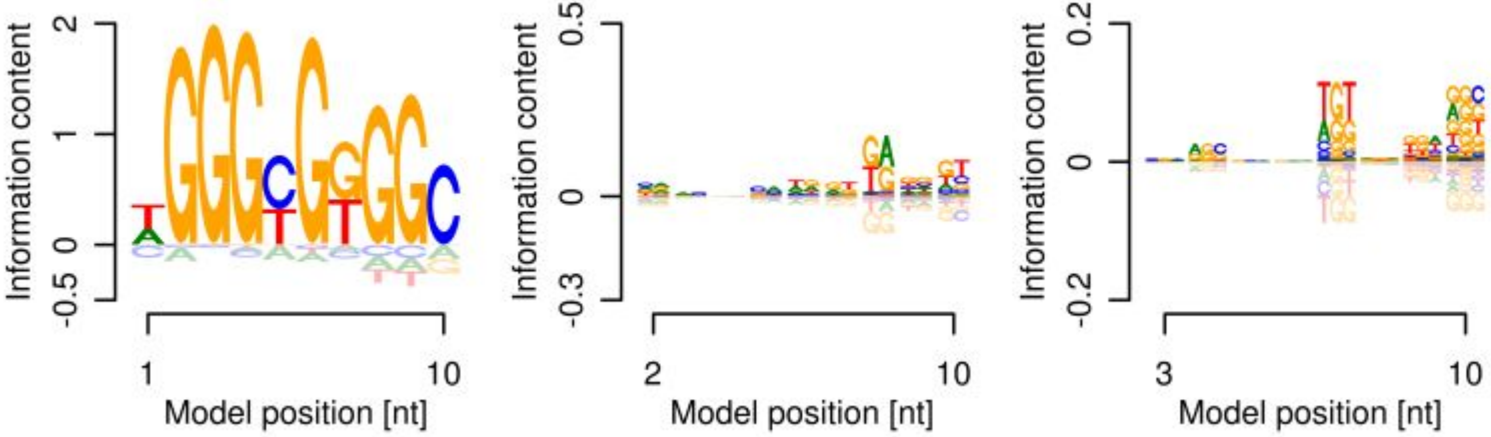
$$p_j^{(0)}(\mathbf{A}) = \frac{n_j(\mathbf{A}) + \alpha_0 p_{\mathrm{bg}}(\mathbf{A})}{N + \alpha_0}$$

$$p_j^{(1)}(\mathbf{A}|\mathbf{T}) = \frac{n_j(\mathbf{TA}) + \alpha_1 p_j^{(0)}(\mathbf{A})}{n_{j-1}(\mathbf{T}) + \alpha_1}$$

$$p_j^{(2)}(\mathbf{A}|\mathbf{CT}) = \frac{n_j(\mathbf{CTA}) + \alpha_2 p_j^{(1)}(\mathbf{A}|\mathbf{T})}{n_{j-1}(\mathbf{CT}) + \alpha_2}$$

$$p_j^{(3)}(\mathbf{A}|\mathbf{GCT}) = \frac{n_j(\mathbf{GCTA}) + \alpha_3 p_j^{(2)}(\mathbf{A}|\mathbf{CT})}{n_{j-1}(\mathbf{GCT}) + \alpha_3}$$

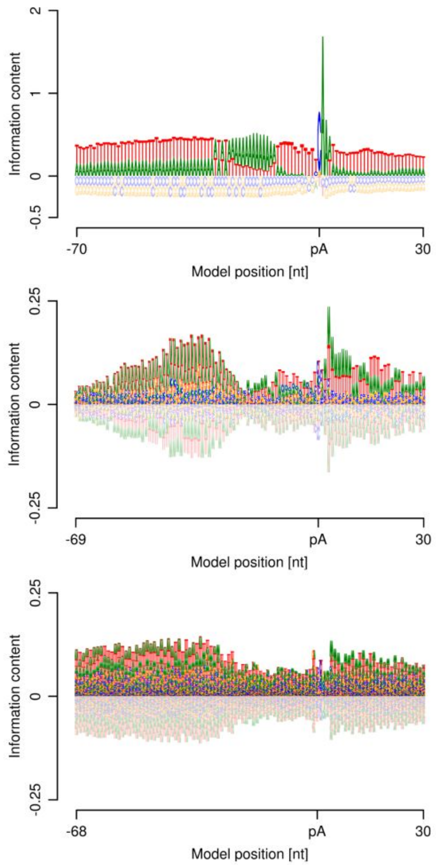
$$\vdots$$



# Complex representation of motifs

**Bayesian Markov Models**

Siebert 2016



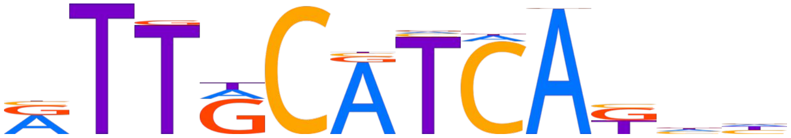
Supports long k-mers

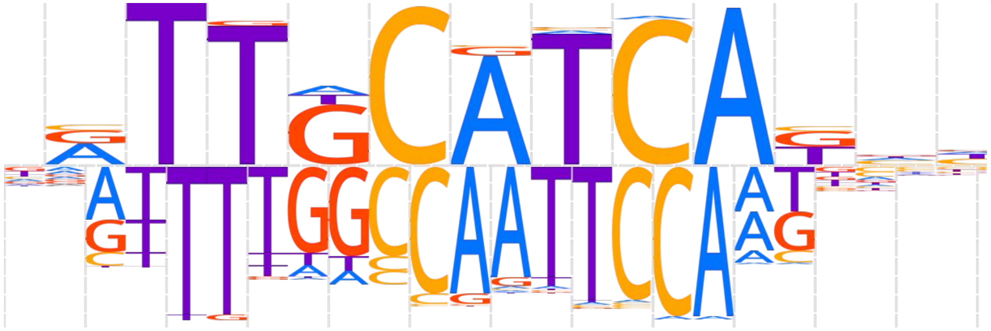
Ideal to visualize dependencies in long sequences, e.g, promoters, poly-A sites

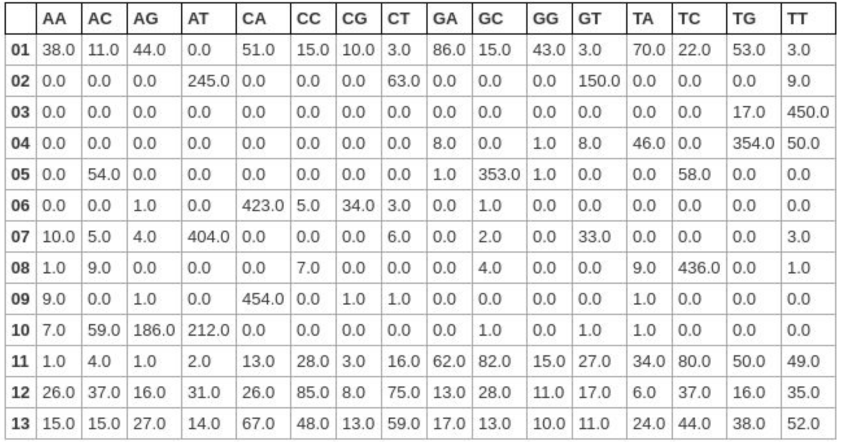
# Complex representation of motifs

**Dinucleotide PWMs**

Kulakovskiy 2018







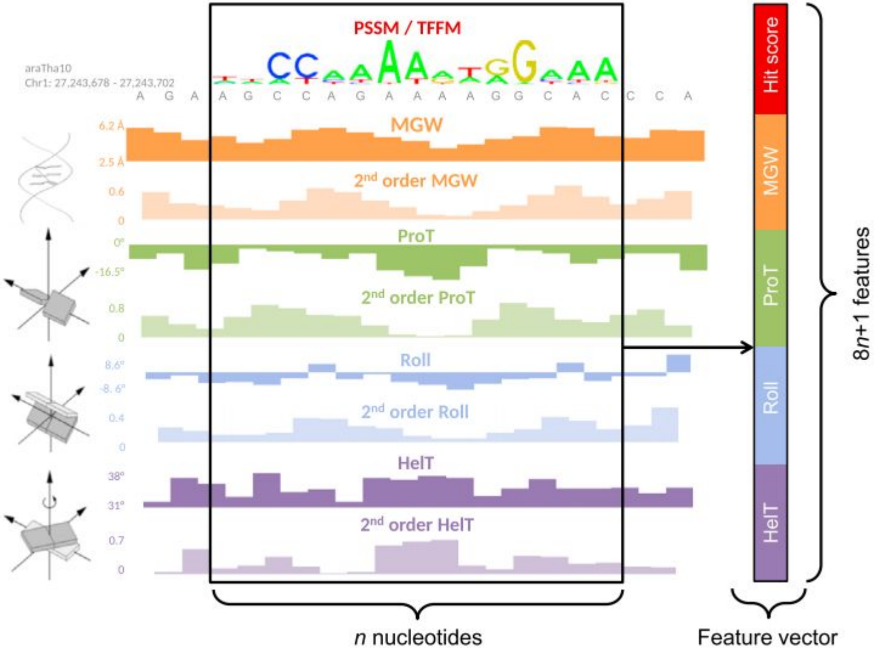
| | AA | AC | AG | AT | CA | CC | CG | CT | GA | GC | GG | GT | TA | TC | TG | TT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 01 | 38.0 | 11.0 | 44.0 | 0.0 | 51.0 | 15.0 | 10.0 | 3.0 | 86.0 | 15.0 | 43.0 | 3.0 | 70.0 | 22.0 | 53.0 | 3.0 |
| 02 | 0.0 | 0.0 | 0.0 | 245.0 | 0.0 | 0.0 | 0.0 | 63.0 | 0.0 | 0.0 | 0.0 | 150.0 | 0.0 | 0.0 | 0.0 | 9.0 |
| 03 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 17.0 | 450.0 |
| 04 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 8.0 | 0.0 | 1.0 | 8.0 | 46.0 | 0.0 | 354.0 | 50.0 |
| 05 | 0.0 | 54.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 353.0 | 1.0 | 0.0 | 0.0 | 58.0 | 0.0 | 0.0 |
| 06 | 0.0 | 0.0 | 1.0 | 0.0 | 423.0 | 5.0 | 34.0 | 3.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 07 | 10.0 | 5.0 | 4.0 | 404.0 | 0.0 | 0.0 | 0.0 | 6.0 | 0.0 | 2.0 | 0.0 | 33.0 | 0.0 | 0.0 | 0.0 | 3.0 |
| 08 | 1.0 | 9.0 | 0.0 | 0.0 | 0.0 | 7.0 | 0.0 | 0.0 | 0.0 | 4.0 | 0.0 | 0.0 | 9.0 | 436.0 | 0.0 | 1.0 |
| 09 | 9.0 | 0.0 | 1.0 | 0.0 | 454.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 10 | 7.0 | 59.0 | 186.0 | 212.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 11 | 1.0 | 4.0 | 1.0 | 2.0 | 13.0 | 28.0 | 3.0 | 16.0 | 62.0 | 82.0 | 15.0 | 27.0 | 34.0 | 80.0 | 50.0 | 49.0 |
| 12 | 26.0 | 37.0 | 16.0 | 31.0 | 26.0 | 85.0 | 8.0 | 75.0 | 13.0 | 28.0 | 11.0 | 17.0 | 6.0 | 37.0 | 16.0 | 35.0 |
| 13 | 15.0 | 15.0 | 27.0 | 14.0 | 67.0 | 48.0 | 13.0 | 59.0 | 17.0 | 13.0 | 10.0 | 11.0 | 24.0 | 44.0 | 38.0 | 52.0 |

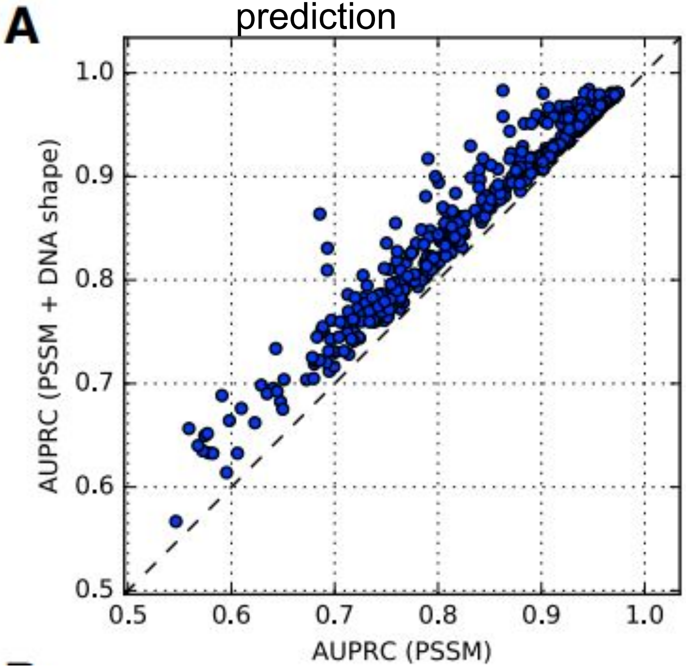
# Complex representation of motifs

**Combining PFMs + DNAshape**

Mathelier 2016

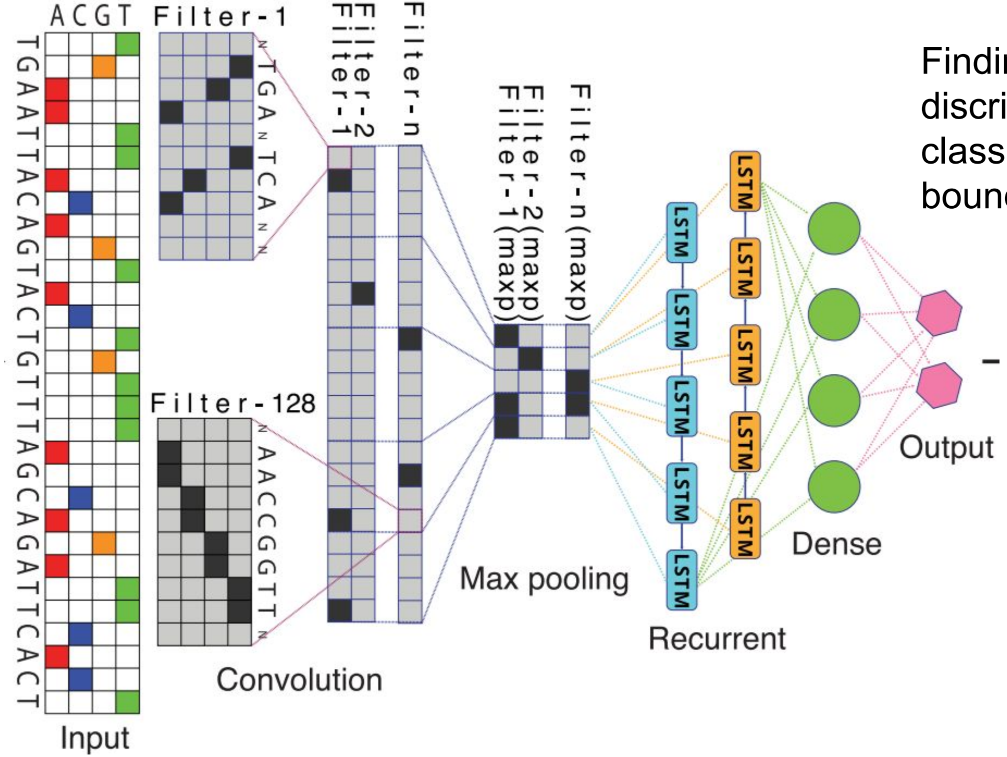
Adding TFBS shape information improves TFBS prediction





# Complex representation of motifs

## Motifs derived from deep learning models

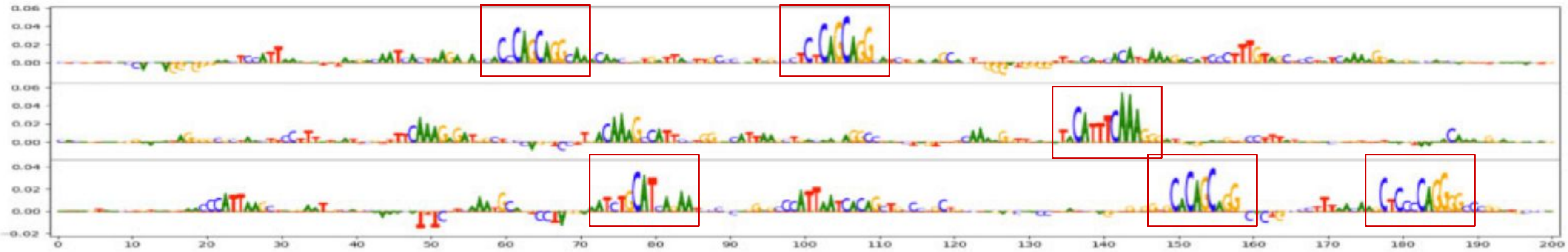


Finding nucleotides that discriminates between classes, e.g., bound/unbound

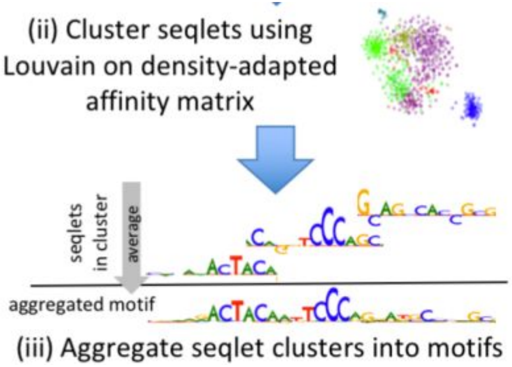
# Complex representation of motifs

**Motifs derived from deep learning or SVM or NMF models**

Shrikumar 2019

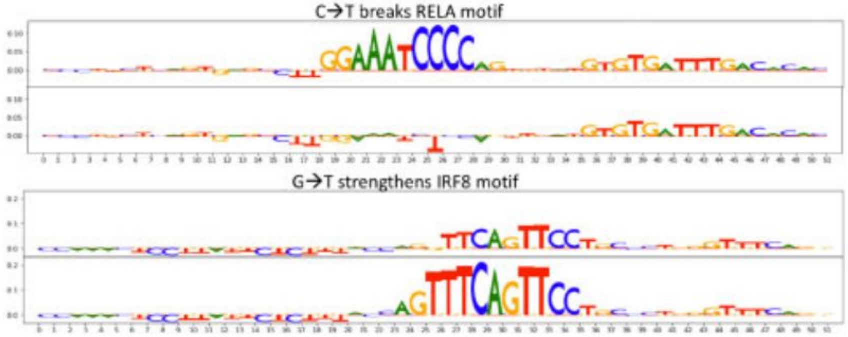


Each nucleotide in the input sequences have a contribution score in the final layer of the NN.



(ii) Cluster seqlets using Louvain on density-adapted affinity matrix

(iii) Aggregate seqlet clusters into motifs



C→T breaks RELA motif

G→T strengthens IRF8 motif

# *Take-home messages*

- Modelling sequences by motifs is an old but still relevant field in bioinformatics (... and it will be always relevant)

- PFMs are still the most used sequence model so far, although more complex alternatives are becoming popular.

- Complex motif representations are not as popular as PFMs, however, they improve TFBSs predictions for particular TF families.
    - No uniform model
    - Many parameters
    - Require large amount of sequences to train

- Methods assessing importance scores allow to detect motif relationship such as motif syntax.

- There is a lot of room for improvement, since many methods were designed to work with the PFMs and not with the more complex models.

# Acknowledgements

- Anthony Mathelier’s group
- Alejandra Medina-Rivera
- JAcques van Helden
- Morgane Thomas-Chollier
- Oriol Fornes

https://mathelierlab.com/