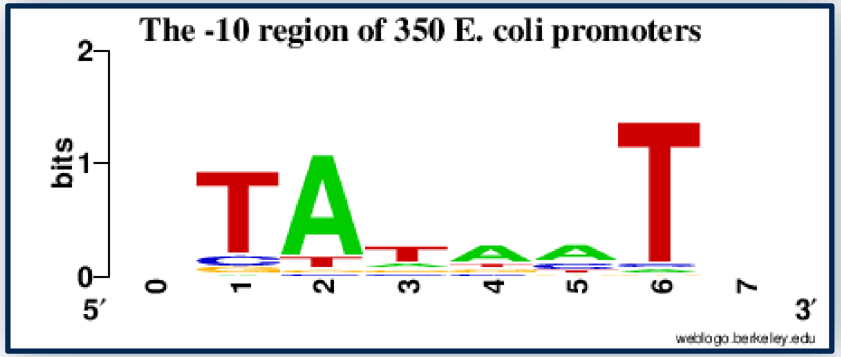
Motif Discovery in DNA and Protein Sequences

Word based and Expectation Maximization based Methods



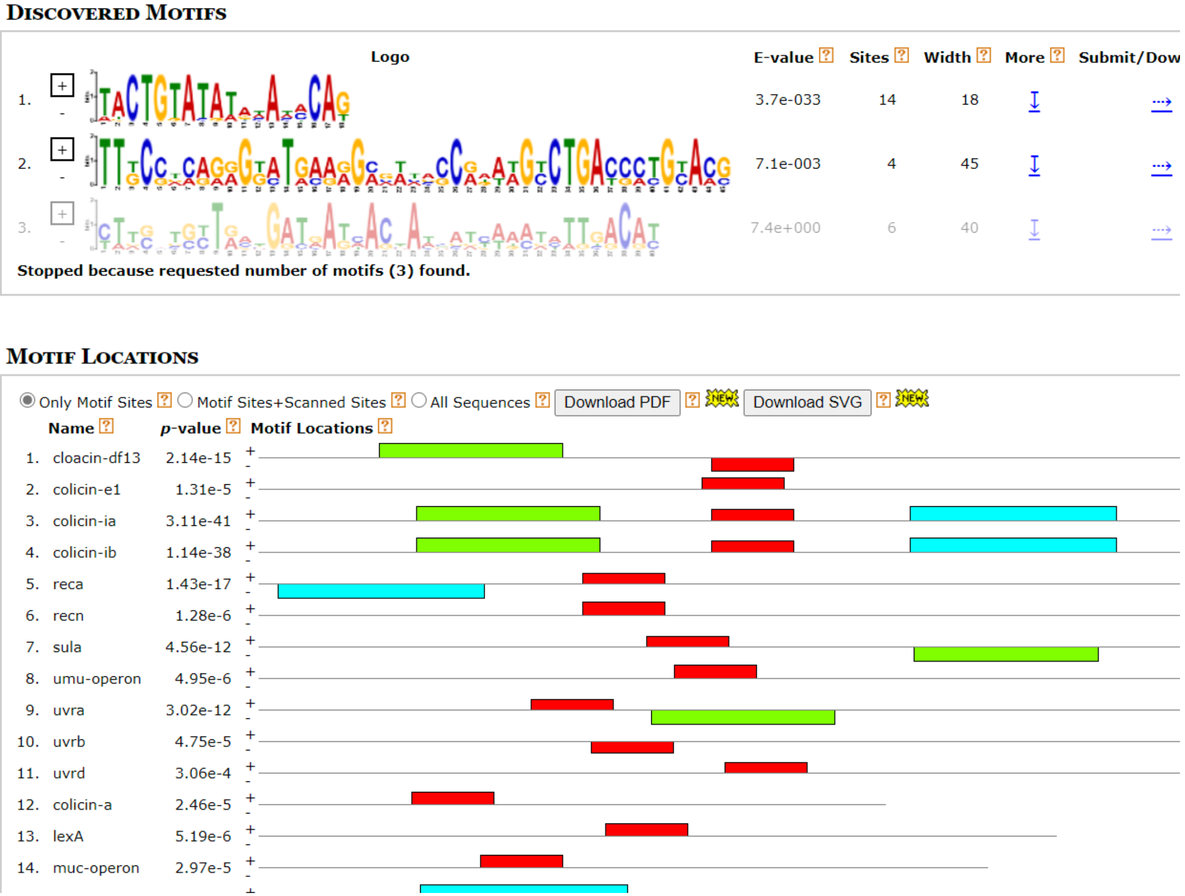
The -10 region of 350 E. coli promoters

Multiple EM for Motif Elicitation

- MEME
  - Is a tool for discovering motifs in a group of related nucleotide or peptide sequences.
  - A MEME motif is a sequence pattern that occurs repeatedly in one or more sequences in the input group.
  - Can be used to discover novel patterns, as it bases its discoveries only on the input sequences, not on any prior knowledge (such as databases of known motifs).
  - MEME motifs allow errors (mutations) at any position in the pattern, but individual MEME motifs may <u>not</u> contain gaps (insertions or deletions).
  - Splits patterns that contain gaps into multiple motifs.
- Motifs may appear in any order, multiple times or not at all in any given sequence.
  - Input
    - a set of unaligned sequences of the same type (peptide or nucleotide) also called training set.

Multiple EM for Motif Elicitation

- For each motif, MEME reports:
  - Discovered motifs;
  - Motif locations;



Discovered Motifs

	Logo	E-value	Sites	Width	More	Submit/Dow
1.		3.7e-033	14	18		
2.		7.1e-003	4	45		
3.		7.4e+000	6	40		

Stopped because requested number of motifs (3) found.

Motif Locations

● Only Motif Sites ○ Motif Sites+Scanned Sites ○ All Sequences Download PDF Download SVG

	Name	*p*-value	Motif Locations
1.	cloacin-df13	2.14e-15	
2.	colicin-e1	1.31e-5	
3.	colicin-ia	3.11e-41	
4.	colicin-ib	1.14e-38	
5.	reca	1.43e-17	
6.	recn	1.28e-6	
7.	sula	4.56e-12	
8.	umu-operon	4.95e-6	
9.	uvra	3.02e-12	
10.	uvrb	4.75e-5	
11.	uvrd	3.06e-4	
12.	colicin-a	2.46e-5	
13.	lexA	5.19e-6	
14.	muc-operon	2.97e-5	

Discovering motifs in a set of Peptides sequences

```
meme filename.fasta -nmotifs 4 -o run1

meme filename.fasta -nmotifs 4 -minw 4 -maxw 10 -o run2
```

oops	One Occurrence Per Sequence
zoops	Zero or One Occurrence Per Sequence MEME
anr	Any Number of Repetitions (This option can also be used to discover repeats within a single sequence)

Allowing repeated motifs

```
meme filename.fasta -nmotifs 4 -minw 4 -maxw 10 -mod anr -o run.anr
```

http://meme-suite.org/doc/meme.html?man_type=web

Word based Motif Discovery with RSAT

- The regulatory sequences analysis tools (RSAT) are a suite of specialized programs for detecting regulatory elements.
  - available at: http://rsat.sb-roscoff.fr/
- The oligo-analysis tool uses an exhaustive approach by scanning all the oligomers of a given size (min and max length can be defined) counting the respective occurrences in a set of sequences. It then uses statistical analysis to detect overrepresented and significant oligonucleotides.

Motif Representations

- Representations of the binding specificity for the Kruppel transcription factor of *Drosophila melanogaster*:

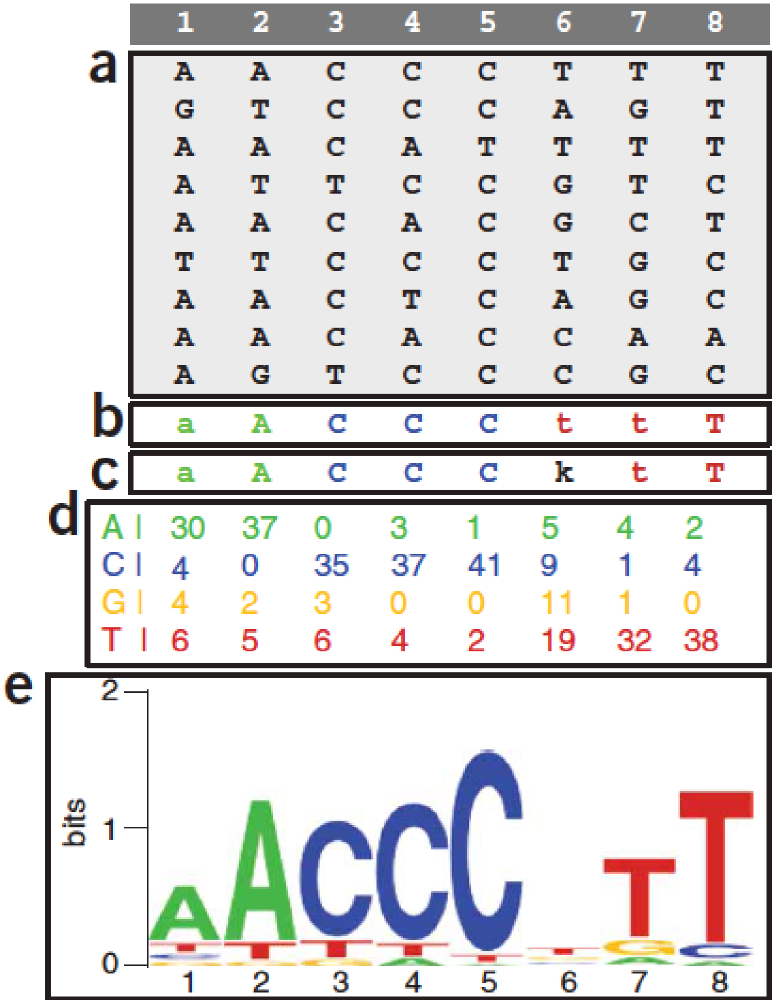
  a) Kruppel site sequences;

  b) Consensus of the above sites;

  c) Degenerate consensus;

  d) Position-specific scoring matrix (PSSM);

  e) Sequence Logo obtained using WebLogo;



	1	2	3	4	5	6	7	8
a	A	A	C	C	C	T	T	T
	G	T	C	C	C	A	G	T
	A	A	C	A	T	T	T	T
	A	T	T	C	C	G	T	C
	A	A	C	A	C	G	C	T
	T	T	C	C	C	T	G	C
	A	A	C	T	C	A	G	C
	A	A	C	A	C	C	A	A
	A	G	T	C	C	C	G	C
b	a	A	C	C	C	t	t	T
c	a	A	C	C	C	k	t	T

d								
A	30	37	0	3	1	5	4	2
C	4	0	35	37	41	9	1	4
G	4	2	3	0	0	11	1	0
T	6	5	6	4	2	19	32	38

Motif Discovery with oligo-analysis tool in DNA sequences

http://rsat.ulb.ac.be/rsat/

Left Panel > Motif Discovery > oligo-analysis

Select:

- Sequence type
- Oligomer Lengths
- Select Background Model

Select GO

In the results select “string-based pattern matching”

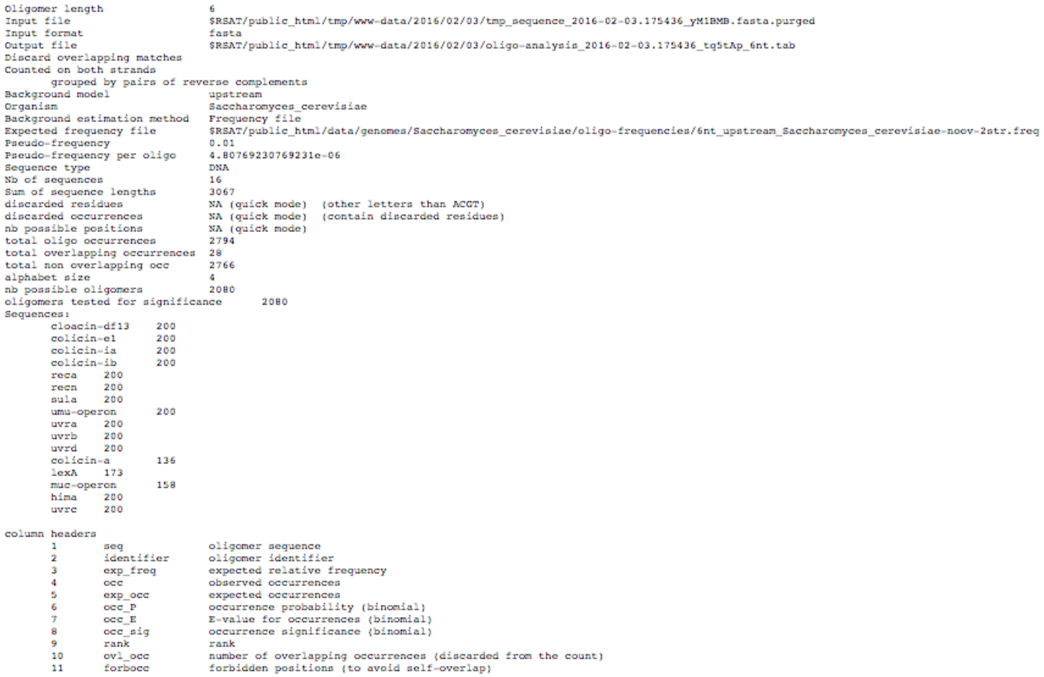
Select GO

Feature Map

Select GO

Results from oligo-analysis

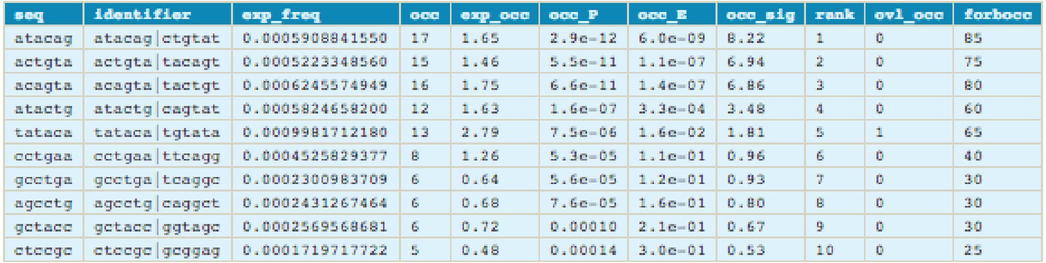
Header with parameters used for the analysis



```
Oligomer length                 6
Input file                      $RSAT/public_html/tmp/www-data/2016/02/03/tmp_sequence_2016-02-03.175436_yM1BMB.fasta.purged
Input format                    fasta
Output file                     $RSAT/public_html/tmp/www-data/2016/02/03/oligo-analysis_2016-02-03.175436_tq5tAp_6nt.tab
Discard overlapping matches
Counted on both strands
       grouped by pairs of reverse complements
Background model                upstream
Organism                        Saccharomyces_cerevisiae
Background estimation method    Frequency file
Expected frequency file         $RSAT/public_html/data/genomes/Saccharomyces_cerevisiae/oligo-frequencies/6nt_upstream_Saccharomyces_cerevisiae-noov-2str.freq
Pseudo-frequency                0.01
Pseudo-frequency per oligo      4.80769230769231e-06
Sequence type                   DNA
Nb of sequences                 16
Sum of sequence lengths         3067
discarded residues              NA (quick mode)  (other letters than ACGT)
discarded occurrences           NA (quick mode)  (contain discarded residues)
nb possible positions           NA (quick mode)
total oligo occurrences         2794
total overlapping occurrences   28
total non overlapping occ       2766
alphabet size                   4
nb possible oligomers           2080
oligomers tested for significance       2080
Sequences:
       cloacin-df13     200
       colicin-e1       200
       colicin-ia       200
       colicin-ib       200
       reca     200
       recn     200
       sula     200
       umu-operon       200
       uvra     200
       uvrb     200
       uvrd     200
       colicin-a        136
       lexA     173
       muc-operon       158
       hima     200
       uvrc     200

column headers
       1        seq             oligomer sequence
       2        identifier      oligomer identifier
       3        exp_freq        expected relative frequency
       4        occ             observed occurrences
       5        exp_occ         expected occurrences
       6        occ_P           occurrence probability (binomial)
       7        occ_E           E-value for occurrences (binomial)
       8        occ_sig         occurrence significance (binomial)
       9        rank            rank
       10       ovl_occ         number of overlapping occurrences (discarded from the count)
       11       forbocc         forbidden positions (to avoid self-overlap)
```

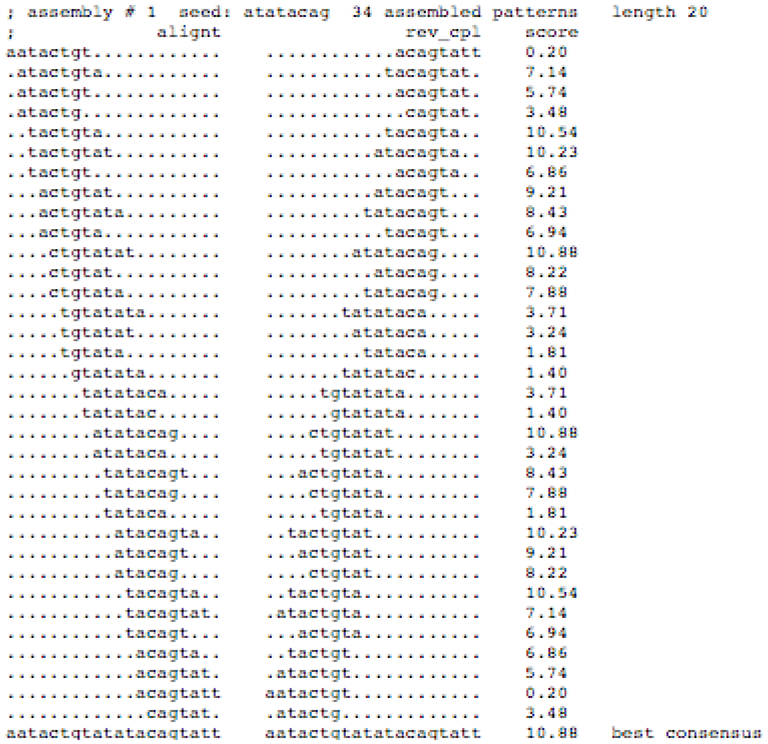
Table with predicted sites. Each row corresponds to a predicted site, defined by its sequence, its coordinates on the input sequence and a series of scores.



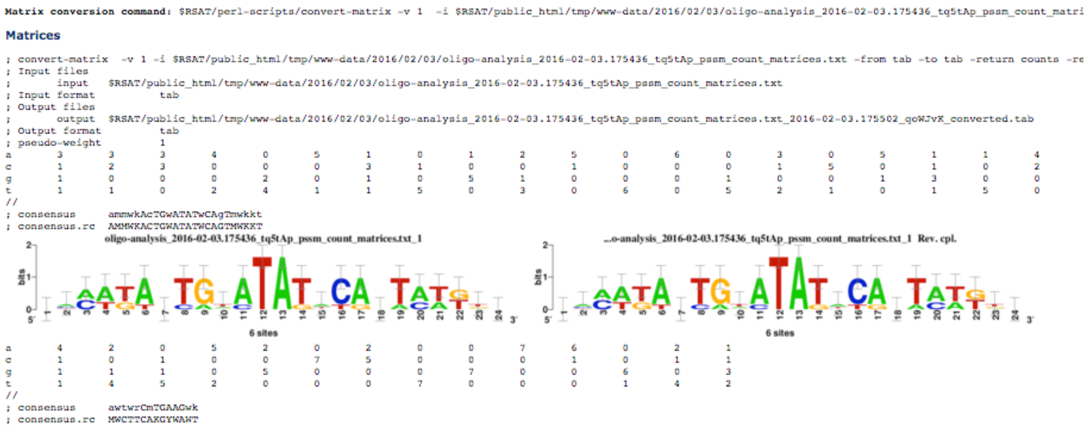
seq	identifier	exp_freq	occ	exp_occ	occ_P	occ_E	occ_sig	rank	ovl_occ	forbocc
atacag	atacag|ctgtat	0.0005908841550	17	1.65	2.9e-12	6.0e-09	8.22	1	0	85
actgta	actgta|tacagt	0.0005223348560	15	1.46	5.5e-11	1.1e-07	6.94	2	0	75
acagta	acagta|tactgt	0.0006245574949	16	1.75	6.6e-11	1.4e-07	6.86	3	0	80
atactg	atactg|cagtat	0.0005824658200	12	1.63	1.6e-07	3.3e-04	3.48	4	0	60
tataca	tataca|tgtata	0.0009981712180	13	2.79	7.5e-06	1.6e-02	1.81	5	1	65
cctgaa	cctgaa|ttcagg	0.0004525829377	8	1.26	5.3e-05	1.1e-01	0.96	6	0	40
gcctga	gcctga|tcaggc	0.0002300983709	6	0.64	5.6e-05	1.2e-01	0.93	7	0	30
agcctg	agcctg|caggct	0.0002431267464	6	0.68	7.6e-05	1.6e-01	0.80	8	0	30
gctacc	gctacc|ggtagc	0.0002569568681	6	0.72	0.00010	2.1e-01	0.67	9	0	30
ctccgc	ctccgc|gcggag	0.0001719717722	5	0.48	0.00014	3.0e-01	0.53	10	0	25

Results from oligo-analysis

Pattern Assembly



```
; assembly # 1  seed: atatacag  34 assembled patterns  length 20
;             alignt                  rev_cpl     score
aatactgt............    ............acagtatt    0.20
.atactgta...........    ...........tacagtat.    7.14
.atactgt............    ............acagtat.    5.74
.atactg.............    .............cagtat.    3.48
..tactgta...........    ...........tacagta..    10.54
..tactgtat..........    ..........atacagta..    10.23
..tactgt............    ............acagta..    6.86
...actgtat..........    ..........atacagt...    9.21
...actgtata.........    .........tatacagt...    8.43
...actgta...........    ...........tacagt...    6.94
....ctgtatat........    ........atatacag....    10.88
....ctgtat..........    ..........atacag....    8.22
....ctgtata.........    .........tatacag....    7.88
.....tgtatata.......    .......tatataca.....    3.71
.....tgtatat........    ........atataca.....    3.24
.....tgtata.........    .........tataca.....    1.81
......gtatata.......    .......tatatac......    1.40
.......tatataca.....    .....tgtatata.......    3.71
.......tatatac......    ......gtatata.......    1.40
........atatacag....    ....ctgtatat........    10.88
........atataca.....    .....tgtatat........    3.24
.........tatacagt...    ...actgtata.........    8.43
.........tatacag....    ....ctgtata.........    7.88
.........tataca.....    .....tgtata.........    1.81
..........atacagta..    ..tactgtat..........    10.23
..........atacagt...    ...actgtat..........    9.21
..........atacag....    ....ctgtat..........    8.22
...........tacagta..    ..tactgta...........    10.54
...........tacagtat.    .atactgta...........    7.14
...........tacagt...    ...actgta...........    6.94
............acagta..    ..tactgt............    6.86
............acagtat.    .atactgt............    5.74
............acagtatt    aatactgt............    0.20
.............cagtat.    .atactg.............    3.48
aatactgtatatacagtatt    aatactgtatatacagtatt    10.88   best consensus
```



Matrix conversion command: $RSAT/perl-scripts/convert-matrix -v 1 -i $RSAT/public_html/tmp/www-data/2016/02/03/oligo-analysis_2016-02-03.175436_tq5tAp_pssm_count_matri

Matrices

```
; convert-matrix  -v 1 -i $RSAT/public_html/tmp/www-data/2016/02/03/oligo-analysis_2016-02-03.175436_tq5tAp_pssm_count_matrices.txt -from tab -to tab -return counts -re
; Input files
;       input   $RSAT/public_html/tmp/www-data/2016/02/03/oligo-analysis_2016-02-03.175436_tq5tAp_pssm_count_matrices.txt
; Input format          tab
; Output files
;       output  $RSAT/public_html/tmp/www-data/2016/02/03/oligo-analysis_2016-02-03.175436_tq5tAp_pssm_count_matrices.txt_2016-02-03.175502_qoWJvK_converted.tab
; Output format         tab
; pseudo-weight         1
a       3       3       3       4       0       5       1       0       1       2       5       0       6       0       3       0       5       1       1       4
c       1       2       3       0       0       0       3       1       0       0       1       0       0       0       1       5       0       1       0       2
g       1       0       0       0       2       0       1       0       5       1       0       0       0       1       0       0       1       3       0       0
t       1       1       0       2       4       1       1       5       0       3       0       6       0       5       2       1       0       1       5       0
//
; consensus     ammwkAcTGwATATwCAgTmwkkt
; consensus.rc  AMMWKACTGWATATWCAGTMWKKT
```

oligo-analysis_2016-02-03.175436_tq5tAp_pssm_count_matrices.txt_1
bits
2
1
0
5′
3′
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
6 sites
...o-analysis_2016-02-03.175436_tq5tAp_pssm_count_matrices.txt_1 Rev. cpl.
6 sites

```
a       4       2       0       5       2       0       2       0       0       7       6       0       2       1
c       1       0       1       0       0       7       5       0       0       0       1       0       1       1
g       1       1       1       0       5       0       0       0       7       0       0       6       0       3
t       1       4       5       2       0       0       0       7       0       0       0       1       4       2
//
; consensus     awtwrCmTGAAGwk
; consensus.rc  MWCTTCAKGYWAWT
```

Additional information on Motifs

- Several high-quality transcription factor binding profile database exist:
  - JASPAR
    - Vertebrate, nematode, insects, plants, fungi, structural classes
  - TRANSFAC
    - eukaryotic **transcription factors**, their experimentally-proven binding sites, consensus binding sequences (positional weight matrices) and regulated genes.
  - CollecTF
    - database of transcription factor binding sites (**TFBS**) in the Bacteria domain.

> http://www.collectf.org/
> Search for TFs
> select: TFS, species and experimental techniques

- Compare results from the different techniques

Motif Scanning

FIMO
Find Individual Motif Occurences

FIMO scans a sequence database for individual matches to each of the motifs provided.

MAST
Motif Alignment & Search Tool

MAST searches sequences for matches to a set of motifs and sorts the sequences by the best combined match to all motifs.

Go to: http://meme-suite.org/tools/fimo

- ○ Select the consensus from oligo-analysis
- ○ Select the species fot the input sequence

Analyze the results. How many sequences have matches?

Motif Discovery in DNA and Protein Sequences

Word based and Expectation Maximization based Methods

Material partially adapted from:
Discovering Novel Sequence Motifs with MEME, Current Protocols in Bioinformatics 2002, TL Bailey
Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules, Nat. Protocols 2008, Turatsinze et al.