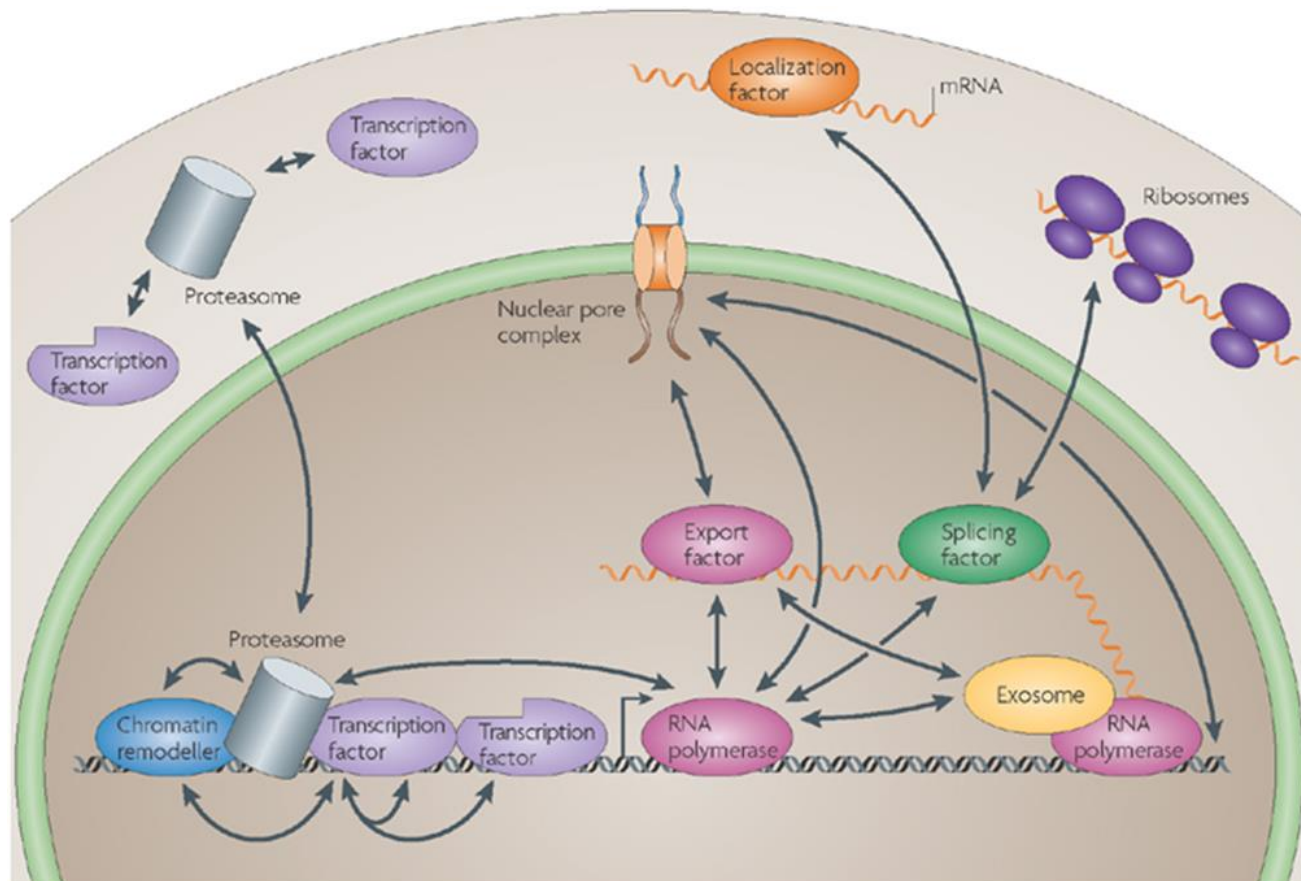


Identification of DNA motifs

Day 2

Gene regulation

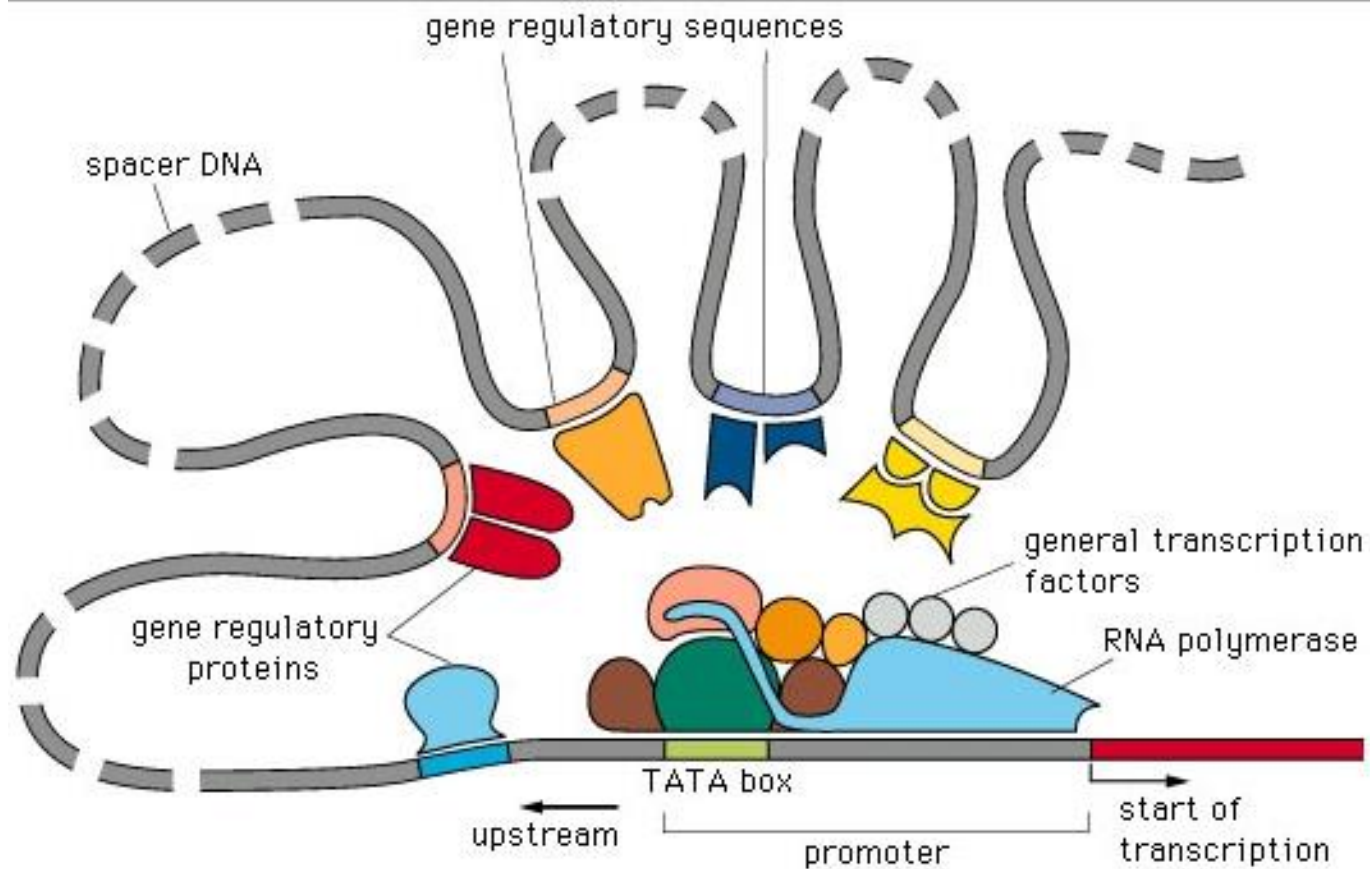


Nature Reviews | **Genetics**

Suzanne Komili & Pamela A. Silver (2008) *Nat. Rev. Genetics* 9:38-48

Gene regulation

Regulation of transcription



http://pbil.univ-lyon1.fr/baobab/article.php3?id_article=278

TRANSFAC® DNA binding sites

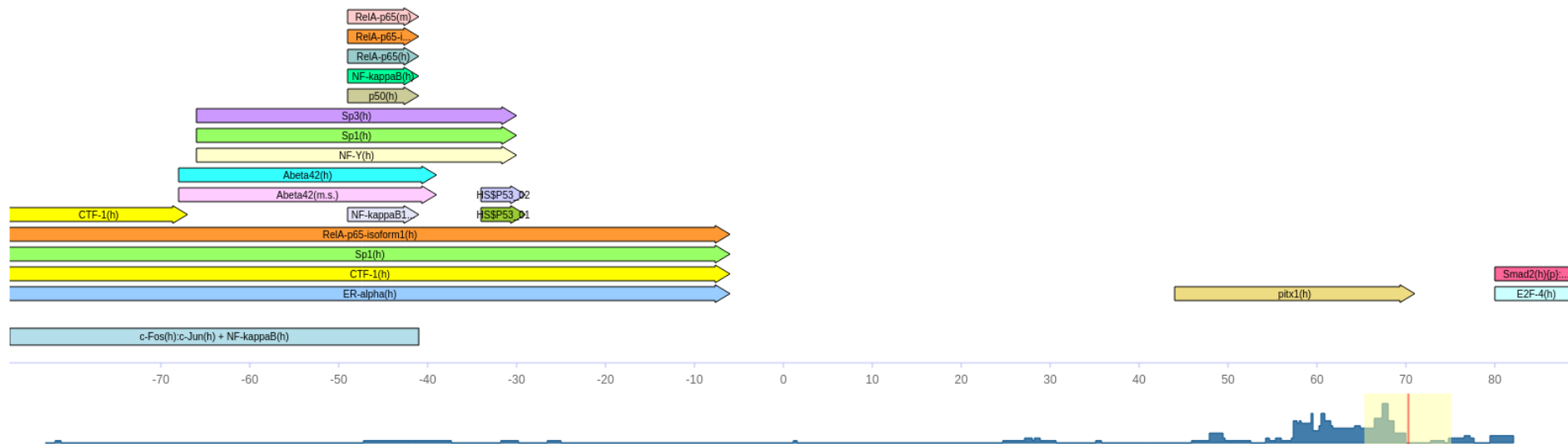
Transcriptional Regulation what is this?

Regulation of TP53 gene expression

Predicted promoter sequences :  Match →

Best supported : PM000860854

All promoters for the gene : PM000636100, PM000860854



Viewing nucleotides -87 to 89

* Note: Only binding sites whose location is relative to the TSS are graphically displayed. Binding sites with an asterisk (*) are not included in the graphical display.

TRANSFAC® DNA sequence motifs

[Table of Contents](#)

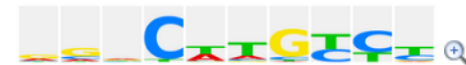
Matrix Overview [what is this?](#)

Consensus sequence logo



Nucleotide position frequency

A	C	G	T	Consensus
25	3	16	2	R
14	0	32	0	G
25	0	21	0	R
2	39	4	1	C
32	2	4	8	A
23	2	2	19	W
3	0	43	0	G
9	15	5	17	N
2	28	9	7	C
5	22	5	14	Y



Nucleotide position frequency

A	C	G	T	Consensus
14	5	22	5	R
7	9	28	2	G
17	5	15	9	N
0	43	0	3	C
19	2	2	23	W
8	4	2	32	T
1	4	39	2	G
0	21	0	25	Y
0	32	0	14	C
2	16	3	25	Y

TRANSFAC® DNA sequence motifs



Aligned sites (46)

Show 50 ▾ entries

Sequence ▴	Transcription factors ▴	Bound Gene ▴	Experimental evidence ▴	Experimental source ▴	References
AAATAGCTC	p53-isoform1(h)	FANCC(h) [details]	CI DM FA FO GS SS IP SE OT	rec(human-E.coli)	36 ↗
AAACAAGCCA	p53-isoform1(h)	IGFBP3(h) [details]	CI DM FA FO GS SS IP SE OT	rec(human-) [more...]	27 ↗ , 44 ↗
AAACATGTCA	p53-isoform1(h)	Fas(m) [details]	CI DM FA FO GS SS IP SE OT	rec(human-Sf9) [more...]	22 ↗
AAACATGTCC	p53-isoform1(h)	FANCC(h) [details]	CI DM FA FO GS SS IP SE OT	rec(human-E.coli)	36 ↗
AAGCAAGCCT	p53-isoform1(h)	Lrdd(m) [details]	CI DM FA FO GS SS IP SE OT	rec(human-) [more...]	19 ↗
AAGCATGTTG	p53-isoform1(h)	IGFBP3(h) [details]	CI DM FA FO GS SS IP SE OT	rec(human-) [more...]	27 ↗ , 44 ↗

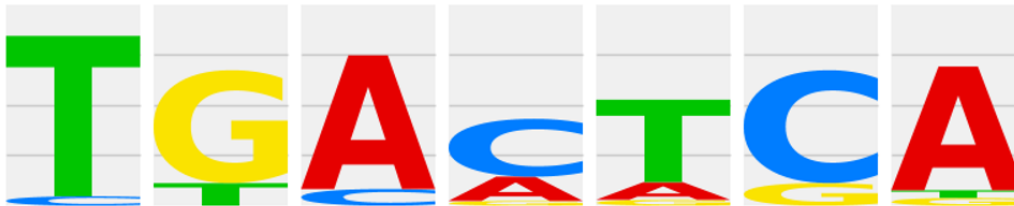
Describing DNA-sequence patterns

Position-specific frequency matrix (PFM) –
the simplest and most often
used DNA-sequence motif model

Position 1234567
Seq1 **TGACTGA**
Seq2 **TGAATCA**
Seq3 **TGAAACA**
Seq4 **TTACTCA**
Seq5 **TGCCTCA**

			1	2	3	4	5	6	7
A	0042105	→	0.0	0.0	0.8	0.4	0.0	0.0	1.0
C	0013040	→	0.0	0.0	0.2	0.6	0.2	0.8	0.0
G	0400010		0.0	0.8	0.0	0.0	0.0	0.2	0.0
T	5100400		1.0	0.2	0.0	0.0	0.8	0.0	0.0
	5555555								

Motif logo of TRANSFAC matrix: V\$AP1_02

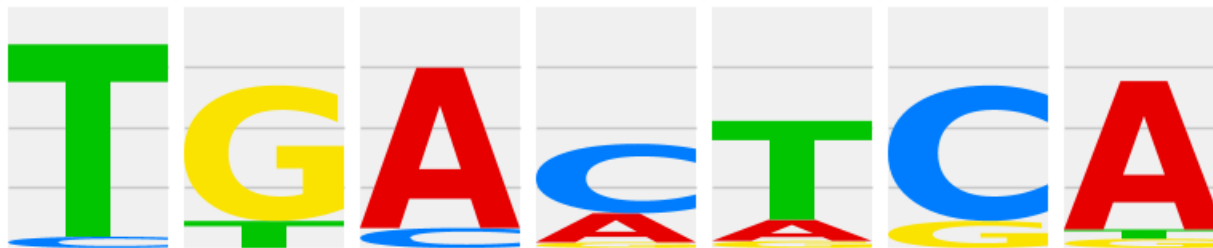


Positional Weight Matrix – binding site scores

Log-odds score of a site: $S_{lod}(x) = \sum_{k=1..L} \log(4 \cdot \frac{n_{ik}+q_i}{N+\sum_i q_i})$

MatchTM score of a site: $S_{Match}(x) = \sum_{k=1..L} I_k \cdot \frac{n_{ik}}{N}$; $I_k = 2 + \sum_i p_{ik} \cdot \log_2 p_{ik}$

	1	2	3	4	5	6	7
A	0.0	0.0	0.8	0.4	0.0	0.0	1.0
C	0.0	0.0	0.2	0.6	0.2	0.8	0.0
G	0.0	0.8	0.0	0.0	0.0	0.2	0.0
T	1.0	0.2	0.0	0.0	0.8	0.0	0.0



Positional Weight Matrix – binding site scores

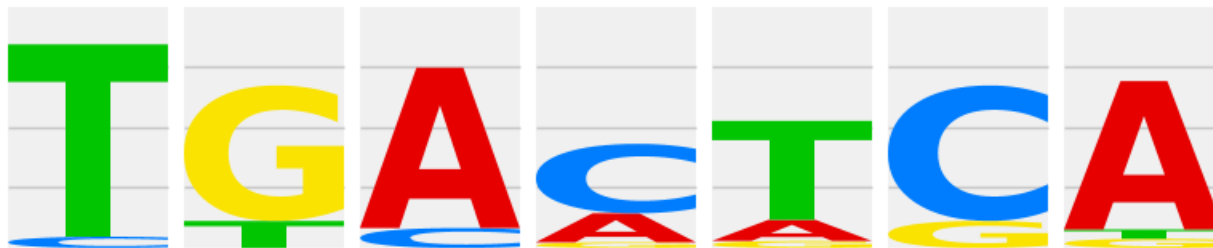
Log-odds score of a site: $S_{lod}(x) = \sum_{k=1..L} \log(4 \cdot \frac{n_{ik}+q_i}{N+\sum_i q_i})$

MatchTM score of a site: $S_{Match}(x) = \sum_{k=1..L} I_k \cdot \frac{n_{ik}}{N}$; $I_k = 2 + \sum_i p_{ik} \cdot \log_2 p_{ik}$

A log-odds score matrix of V\$AP1_02

A	-4.25	-4.25	1.77	0.14	-0.55	-4.25	1.77
C	-1.93	-4.25	-1.08	1.37	-4.25	1.68	-4.25
G	-4.25	1.68	-4.25	-1.93	-1.93	-0.55	-1.93
T	1.86	-0.55	-4.25	-4.25	1.58	-4.25	-1.93

Score of the consensus site: 11.71



DNA binding site prediction

- Score thresholds
- Profiles
- Site p -values

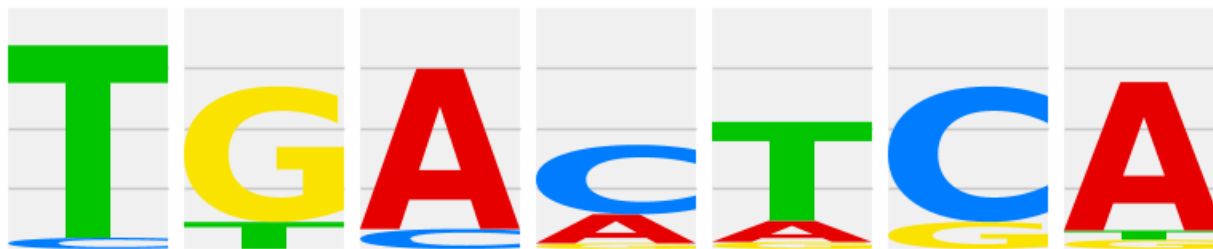
Sequence scan for binding sites

...ACGCGACGGTACT**TGACTCA**ACGTACGAACTT...

.

A	-4.25	-4.25	1.77	0.14	-0.55	-4.25	1.77
C	-1.93	-4.25	-1.08	1.37	-4.25	1.68	-4.25
G	-4.25	1.68	-4.25	-1.93	-1.93	-0.55	-1.93
T	1.86	-0.55	-4.25	-4.25	1.58	-4.25	-1.93

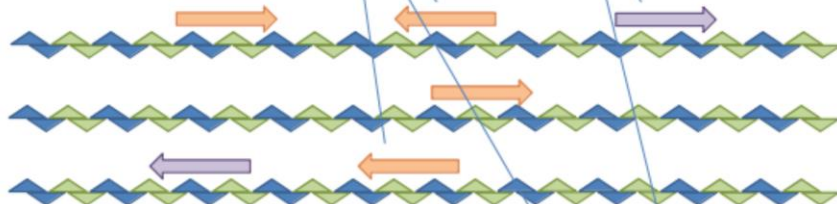
Score of the consensus site: 11.71



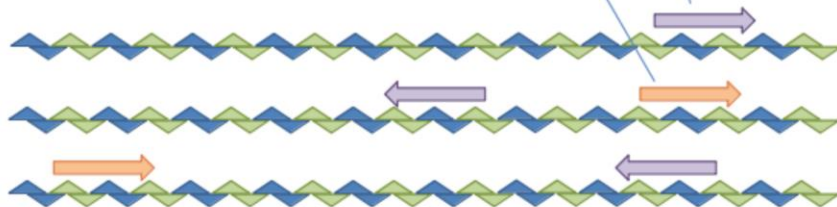
Motif enrichment analysis

TTC_ATT AAC_TG

Yes sequences



No sequences



ID	Adj. site FE	Site P-value	Site FDR	%Yes seq	Adj. seq FE	Seq FDR
V\$NFKAPPAB_01	3.24528	5.5515E-11	2.0124E-8	22.22222	3.0466	4.4505E-6
V\$CREL_Q6	3.03578	1.7049E-7	9.1557E-6	14.14141	3.03578	3.7501E-5
V\$CREL_01	2.64766	2.5786E-9	3.3228E-7	26.26263	2.47123	4.4505E-6
V\$NFKB_Q6_01	2.62117	5.297E-8	3.2569E-6	20.20202	2.98341	4.4505E-6
V\$P50_Q6	2.55339	7.782E-7	3.224E-5	16.16162	2.82629	3.7501E-5
V\$NFKAPPAB65_01	2.48902	7.1794E-7	3.0618E-5	16.16162	2.55339	1.0081E-4
V\$RELA_Q6	2.41035	1.0443E-10	3.0286E-8	11.11111	2.07655	0.00261
V\$RELB_P52_01	2.37496	2.4646E-6	6.9853E-5	14.14141	2.197	8.5918E-4
V\$NFKB_C	2.32939	7.897E-6	1.7091E-4	11.11111	2.32939	8.5918E-4
V\$NFKB_Q6	2.27382	1.5376E-8	1.2386E-6	10.10101	1.86388	0.00844
V\$P50RELAP65_Q5_01	2.24067	8.4394E-6	1.7735E-4	13.13131	2.24067	8.5918E-4
V\$P50P50_Q3	1.92704	2.7499E-9	3.3228E-7	33.33333	1.38401	8.5918E-4
V\$EWSR1FLI1_01	1.85679	5.1794E-9	5.0068E-7	100	0.65924	1
V\$POU3F2_01	1.52795	2.5192E-8	1.9051E-6	100	0.66341	1
V\$FOXK1_04	1.52449	1.4113E-4	0.00146	100	0.65924	1
V\$AP2_Q3	1.50708	9.0938E-5	0.0011	18.18182	1.46654	0.0096
V\$ELF5_01	1.48754	2.6172E-5	4.5722E-4	100	0.66132	1
V\$FXR1R1_Q6	1.33904	8.971E-4	0.00447	100	0.65924	1
V\$IRF7_Q3_01	1.30236	3.9453E-4	0.00287	18.18182	1.17456	0.05465
V\$SFPI1_04	1.30236	3.9453E-4	0.00287	100	0.66132	1

Motif enrichment analysis

- Use cases:
 - Find transcription factors for co-regulated gene set
 - Identify motifs “co-enriched” in ChIP-seq regions

- Which binding sites are significantly enriched in the Yes sequences?
- Site optimization tool
- Automatic threshold optimization
- One-sided binomial test for significant enrichment of sites

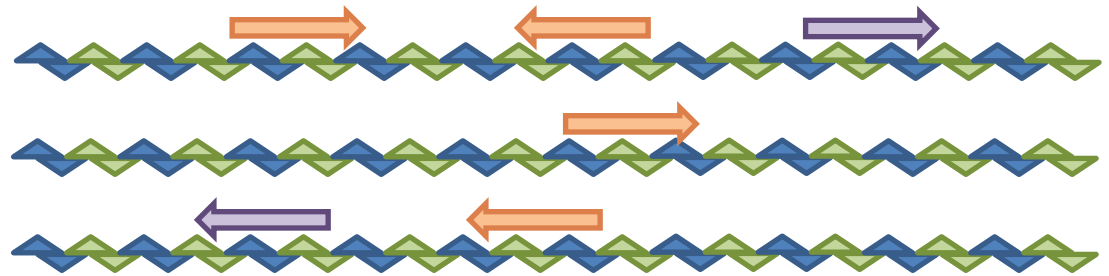
$$P(k \geq n) = \sum_{k=n}^N \binom{N}{k} \cdot p^k \cdot (1-p)^{N-k}$$

$$p = \frac{\#Seq_{Yes}}{\#Seq_{Yes} + \#Seq_{No}}$$

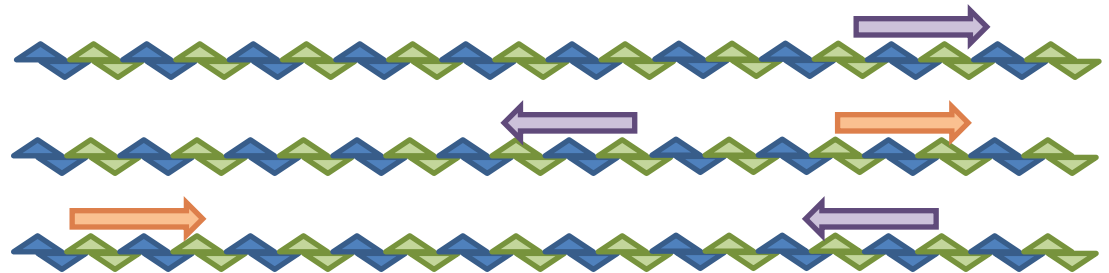
$$n = \#Sites_{Yes}$$

$$N = \#Sites_{Yes} + \#Sites_{No}$$

Yes sequences



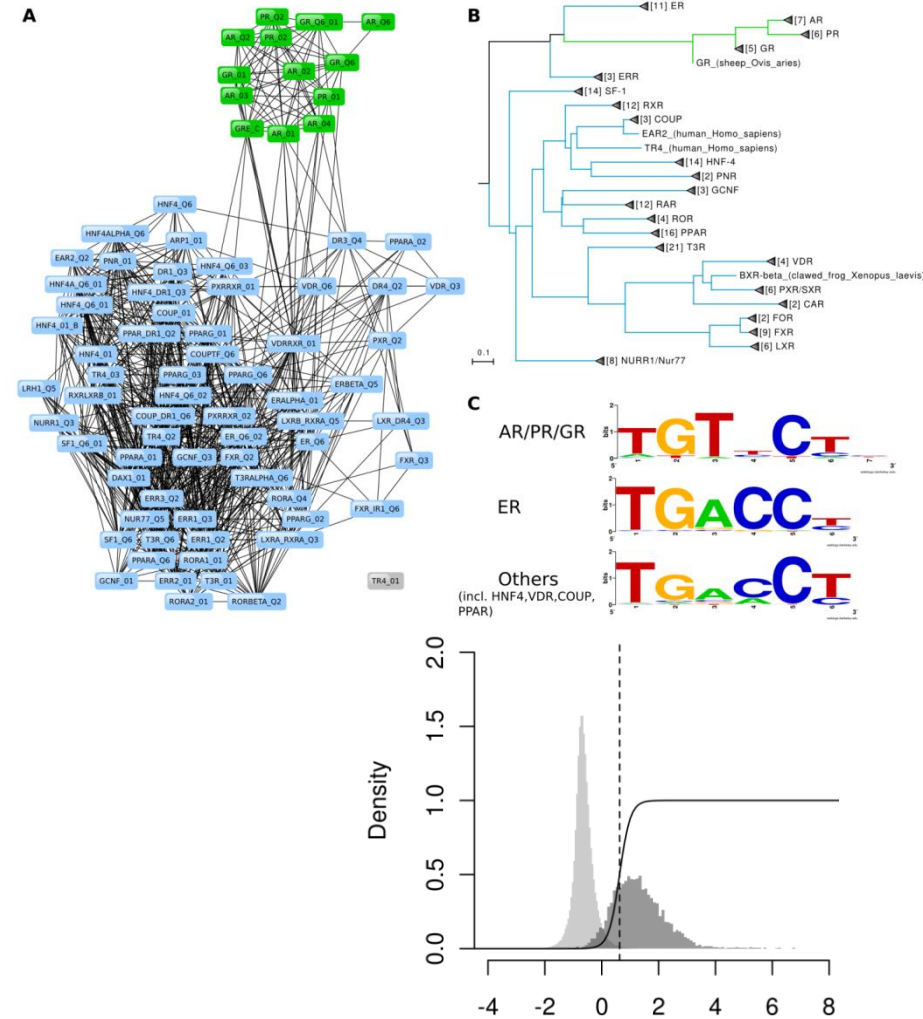
No sequences



Motif enrichment analysis

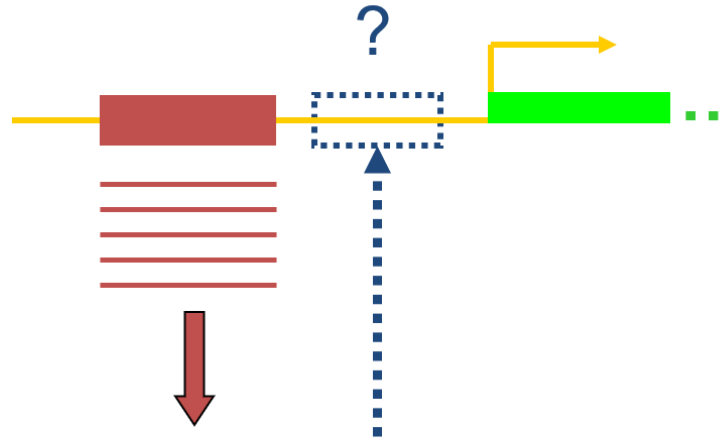
Improved methods in the geneXplain platform

- MEALR: motif enrichment analysis by logistic regression
- Sparse logistic regression to select a subset of motifs from a possibly large and redundant motif library
- $x_{ik} = \log(\frac{1}{L_i} \sum_{w_i} \exp(S_w))$
- $S_w :=$ Log-odds score of window w
- Sequences $i = 1 \dots N$ with classes $\{0,1\}$ and lengths L_i
- Motifs $k = 1 \dots M$
- For enrichment analysis, rank motifs by LR coefficient from highest to smallest.



[Stegmaier et al. \(PLoS CB, 2013\)](#)

Search for new TF binding sites with PWMs

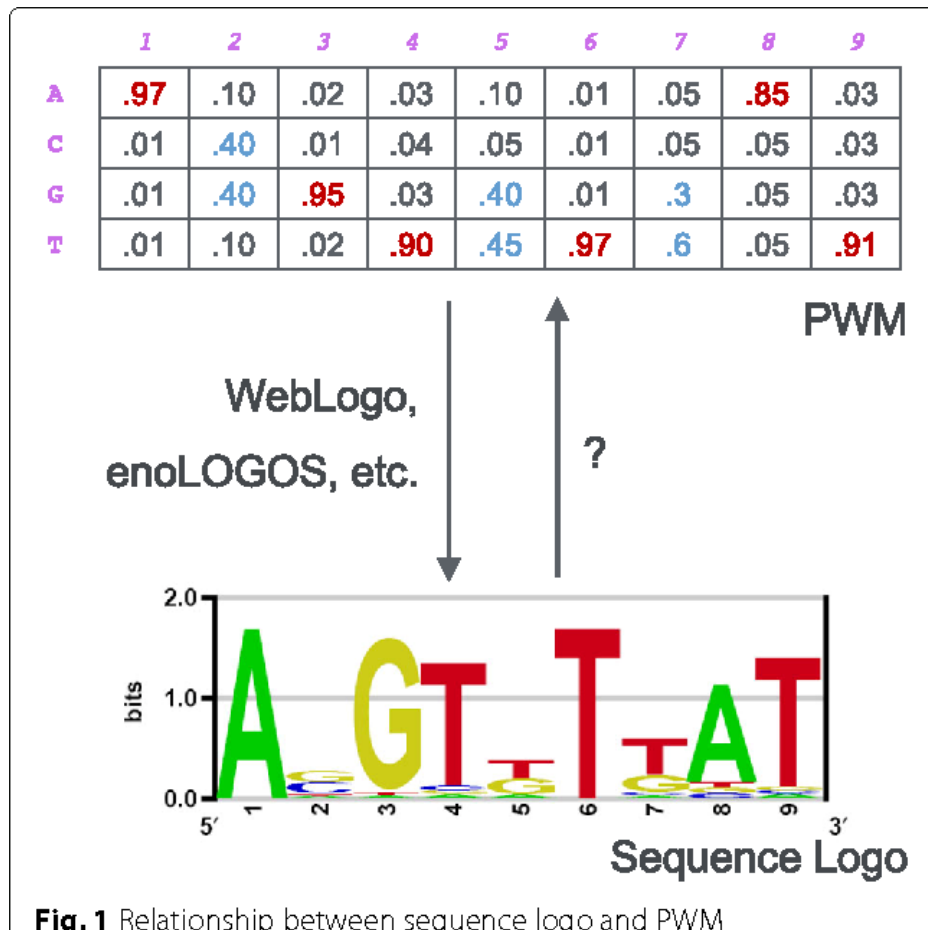


A	9	2	1	0	1	0	0	0	0	1	15	13	13	7
C	8	3	1	1	13	3	29	0	22	8	9	1	4	8
G	4	2	2	2	15	26	0	29	7	17	3	7	9	8
T	8	22	25	26	0	0	0	0	0	3	2	8	3	6
	N	T	T	T	S	G	C	G	C	S	M	D	R	N

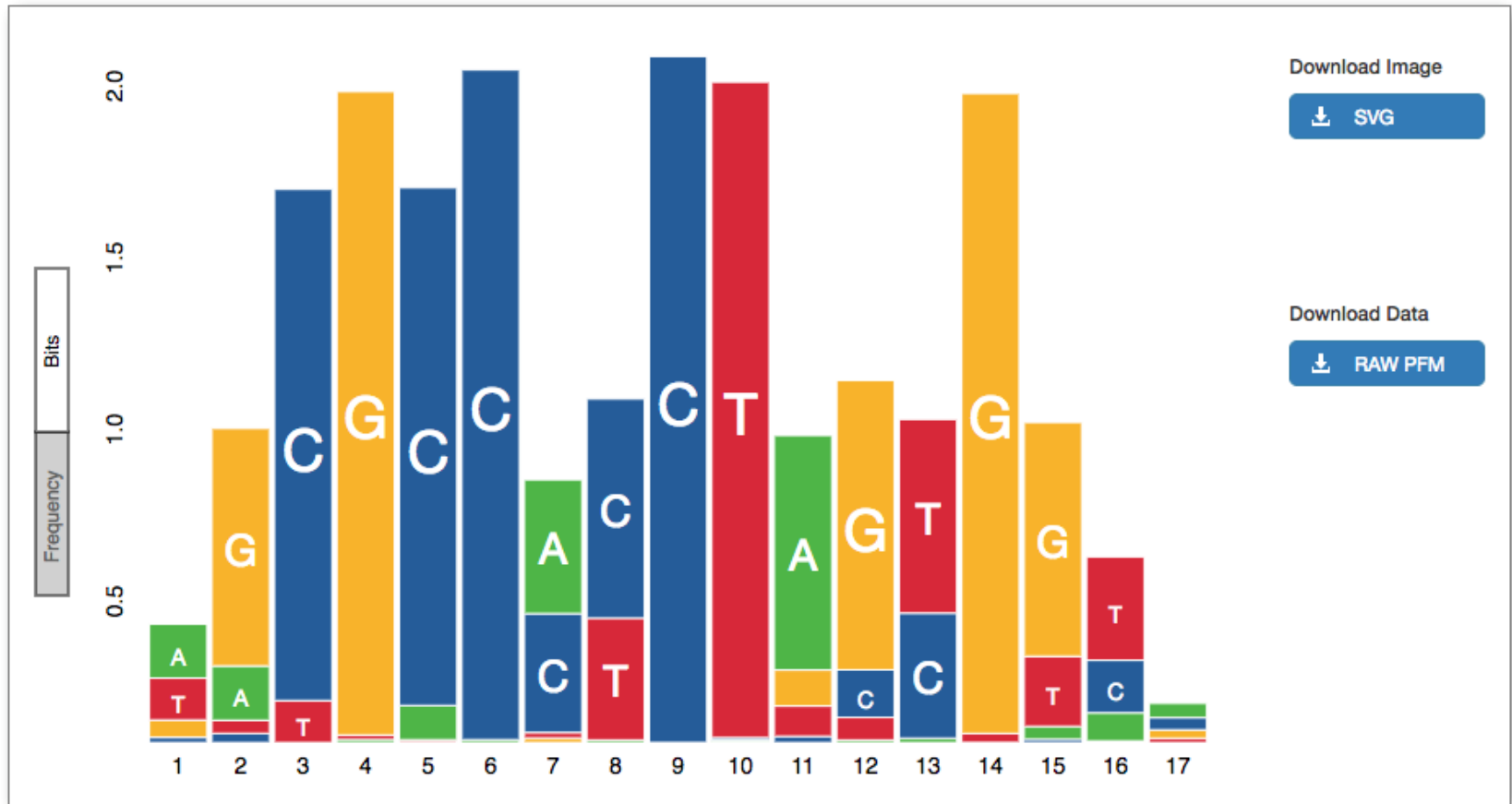
$$q = \frac{\sum_{i=1}^l I(i) f(b_i, i) - \sum_{i=1}^l I(i) f^{\min}(i)}{\sum_{i=1}^l I(i) f^{\max}(i)} \quad (1)$$

$$I(i) = \sum_{b \in \{A, T, G, C\}} f(b, i) \ln(4 f(b, i)) \quad (2)$$

Relationship between Sequence Logo and PWM



New Ensembl motif features



<http://www.ensembl.info/2018/10/15/new-ensembl-motif-features/>

ChIPMunk: fast and efficient motif discovery tool

ChIPMunk: fast and efficient motif discovery tool, reborn and running

ChIPMunk homepage @ autosome.ru

ChIPMunk DNA&RNA motif discovery tool now comes in a single package with **diChIPMunk**, ready to process your ChIP-Seq, HT-SELEX, DNase footprints & similar data, including sequence data on RNA-binding proteins (e.g. PAR-CLIP or CLIP-Seq).

Our sequence-crunching rodents are now accompanied by SPRY-SARUS motif scanner, to apply discovered PWMs to look for motif hits in given sequence sets.

You may also check [MACRO-APE](#) and [PERFECTOS-APE](#) web tools, which are also useful for downstream analysis involving ChIPMunk results.



Need a demo? Try online web-interface for [ChIPMunk](#) and [diChIPMunk](#). Need a simple motif scanner? Try the newest [SARUS](#).

ChIPMunk downloads

[chipmunk.jar](#) ChIPMunk latest (v8) complied classes

[userguide.pdf](#) ChIPMunk detailed user guide

Additional downloads

[chipmunk_v8_src.jar](#) ChIPMunk v8 java sources

[chipmunk_v8.jar](#) ChIPMunk stable (v8, currently the same as above) complied classes

[chipmunk_peaksample.zip](#) ChIPMunk peak fasta examples

[chipmunk_scripts.zip](#) ChIPMunk supporting scripts (ruby)

Please, use the latest versions provided at this page.

Citing ChIPMunk

Deep and wide digging for binding motifs in ChIP-Seq data. Kulakovskiy *et al.*, 2010, [PubMed](#)

From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. Kulakovskiy *et al.*, 2013, [PubMed](#)

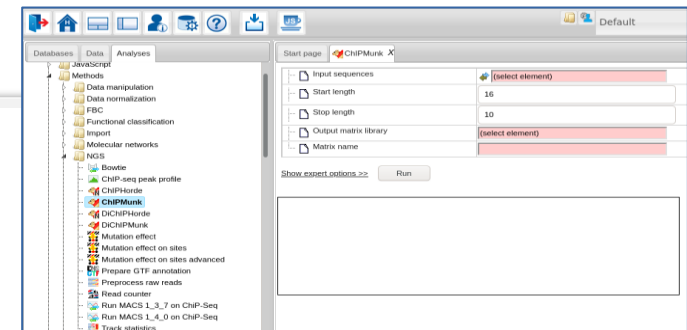
Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. Levitsky *et al.*, 2014, [PubMed](#)

Contacts

In case of any questions don't hesitate to contact [\[ivan-dot-kulakovskiy-at-gmail-dot-com\]](mailto:ivan-dot-kulakovskiy-at-gmail-dot-com).

This software is maintained by Ilya Vorontsov and Ivan Kulakovskiy.

<http://autosome.ru/ChIPMunk/>



Binding site prediction & Identification of enriched motifs

Start page Site search on track X

Track	☐ (select element)
Sequences source	Ensembl 100.38 Human (hg38)
Profile	...enriched motifs_TRANSFAC(R))/Profile
Output name	(select element)

Databases Data Analyses

- Training 290414 P9
- Training_platform
 - Data
 - 1_Introduction
 - 2_Motifs
 - 3_Factors
 - 4_Enhancers
 - 5_Pathways
 - 6_Networks
 - 7_Targets
 - 8_Biomarkers
 - Training_Input_Data
 - 2_Motifs_Input
 - BRCA_gene
 - DRIMust_Encode_NFKB_TNFA_Gm10847_Top500.fasta
 - GSM558469_E2F1_hg19_filtered_chr1
 - GXP_Encode_Gm10847_NfkbTnfa_Top500.narrowPeak
 - UpRegGeneList

Start page Identify enriched motifs ... X

Identify enriched motifs in promoters (TRANSFAC(R))

Input Yes gene set	...t_Data/2_Motifs_Input/UpRegGeneList
Input No gene set	...ata/Housekeeping genes (Human) 300
Annotation source	databases/EnsemblHuman100/Data/gen...
Profile	...ata/profiles/vertebrate_human_p0.001
Species	Human (Homo sapiens)
Filter by TFBS enrichment fold	1.0
Start promoter	-1000
End promoter	100
Allow big input	☐
Result folder	...neList (enriched motifs_TRANSFAC(R))

Run workflow Edit workflow

Hands-on training

- **ChIPMunk**
- **Binding site prediction**