Regulatory Sequence Analysis

Assessing the quality of position-specific scoring matrices

Jacques van Helden

Jacques.van-Helden@univ-amu.fr
Aix-Marseille Université, France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univmed.fr/

FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)
http://www.bigre.ulb.ac.be/

Bioinformatique des Génomes et des Réseaux (BiGRe)
Université Libre de Bruxelles
http://www.bigre.ulb.ac.be/


Nicolas Simonis
Postdoc


Karoline Faust
PhD student


Jean Valéry Turatsinze
PhD student


Jacques van Helden
Chargé de cours


Raphaël Leplae
Postdoc


Ariane Toussaint
Professor


Gipsi Lima
Postdoc


Didier Gonze
Premier assistant


Myriam Loubriat
Secretary


Olivier Sand
Ex-Postdoc


Matthieu Defrance
X-Postdoc


Morgane Thomas-Chollier
Ex-PhD student+postdoc


Sylvain Brohée
Ex-PhD student


Rekin's Janky
Ex-PhD student

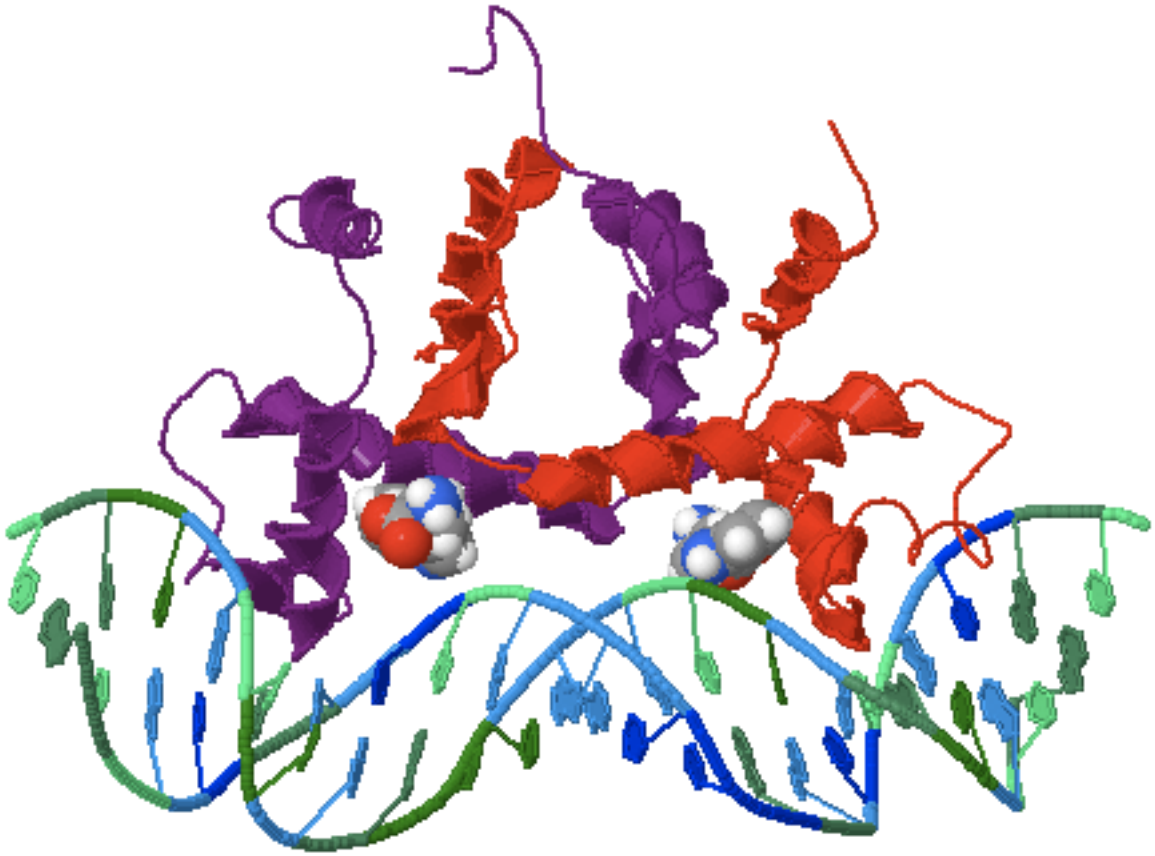

Eric Vervisch
Ex-Research fellow

- *Development and application of bioinformatics methods for the analysis of genome function, regulation and evolution.*
- Regulatory sequences
  - Pattern discovery algorithms
    - *Olivier Sand (Postdoc), Matthieu Defrance (Postdoc), Maud Vidick (Master thesis)*
  - Evolution of cis-acting elements in Bacteria
    - *Rekin's Janky (PhD student)*
  - Regulation of development in Drosophila
    - *Jean Valéry Turatsinze (PhD student)*
  - Hox regulation in Vertebrates
    - *Morgane Thomas-Chollier (PhD student)*
  - Work flows on transcriptional regulation
    - *Olivier Sand (Postdoc), Eric Vervisch (Research fellow)*
- Molecular networks
  - Analysis of regulatory networks
    - *Rekin's Janky (PhD student), Sylvain Brohée (PhD student)*
  - Interactions between membrane-associated proteins
    - *Sylvain Brohée (PhD student)*
  - Inference of metabolic pathways
    - *Karoline Faust (PhD student)*
  - Host-virus interaction networks
    - *Nicolas Simonis (Postdoc)*
- Mobile genetic elements in prokaryotes
  - *Raphaël Leplae (Postdoc), Gipsi Lima (PhD student), Ariane Toussaint (Professor)*
- Modelling of dynamical systems
  - *Didier Gonze (Premier assistant)*

Detailed description

- Detailed description of the ptoram matrix-quality
  - Medina-Rivera A, Abreu-Goodger C, Thomas-Chollier M, Salgado H, Collado-Vides J, van Helden J. 2011. Theoretical and empirical quality assessment of transcription factor-binding motifs. Nucleic Acids Res 39(3): 808-824.
- Web tool available at Regulatory Sequence Analysis Tools (RSAT)
  - http://rsat.bigre.ulb.ac.be/rsat/

NMR study of TrpR



- Zhang et al (1994). The solution structures of the trp repressor-operator DNA complex. J Mol Biol 238:592-614. PDB entry **1rcs.**

TrpR binding sites and motif in Escherichia coli K12

(a) Annotated TrpR binding sites

Site ID																			Target Operon
ECK120012644	G	T	A	C	T	A	G	T	T	T	G	A	T	G	G	T	A	T	G	aroL-yaiA-aroM
ECK120012187	G	T	A	C	T	A	G	T	T	T	G	A	T	G	G	T	A	T	G	aroL-yaiA-aroM
ECK120012179	G	A	A	C	T	A	G	T	T	A	A	C	T	A	G	T	A	C	G	trpLEDCBA
ECK120012892	G	A	A	C	T	A	G	T	T	A	A	C	T	A	G	T	A	C	G	trpLEDCBA
ECK120012181	G	A	A	C	T	A	G	T	T	A	A	C	T	A	G	T	A	C	G	trpLEDCBA
ECK120012636	G	T	A	C	T	A	G	A	G	A	A	C	T	A	G	T	G	C	A	aroH
ECK120012183	G	T	A	C	T	A	G	A	G	A	A	C	T	A	G	T	G	C	A	aroH
ECK120012185	G	T	A	C	T	C	G	T	G	T	A	C	T	G	G	T	A	C	A	mtr
ECK120012979	G	T	A	C	T	C	G	T	G	T	A	C	T	G	G	T	A	C	A	mtr
ECK120012894	G	T	A	C	T	C	T	T	T	A	G	C	G	A	G	T	A	C	A	trpR

(b) Position-specific scoring matrix

A	0	3	**10**	0	0	7	0	2	0	6	7	2	0	6	0	0	8	0	5
T	0	7	0	0	**10**	0	1	8	6	4	0	0	9	0	0	**10**	0	2	0
C	0	0	0	**10**	0	3	0	0	0	0	0	8	0	0	0	0	0	8	0
G	**10**	0	0	0	0	0	9	0	4	0	3	0	1	4	**10**	0	2	0	5

(c) Consensus

G	w	A	C	T	m	G	t	k	w	r	C	t	r	G	T	r	C	r

(d) Sequence logo

Background models - Drosophila non-coding upstream sequences

Drosophila upstream sequences
Bernoulli model (nucleotide frequencies)

a	c	g	t
0.29515	0.20568	0.20539	0.29378

Drosophila upstream sequences
Markov order 1

pr\suf	a	c	g	t
a	0.35787	0.17609	0.18046	0.28557
c	0.33195	0.21597	0.19577	0.25632
g	0.26293	0.26993	0.21370	0.25344
t	0.22881	0.18378	0.23121	0.35620

Drosophila upstream sequences
Markov order 2

pr\suf	a	c	g	t
aa	0.37291	0.16877	0.16923	0.28909
ac	0.35709	0.18621	0.17167	0.28504
ag	0.26165	0.27237	0.18363	0.28236
at	0.26257	0.16673	0.20870	0.36201
ca	0.33684	0.19795	0.20584	0.25936
cc	0.35835	0.22516	0.19563	0.22085
cg	0.30620	0.25800	0.21443	0.22137
ct	0.19334	0.20958	0.26335	0.33373
ga	0.36047	0.17482	0.20549	0.25922
gc	0.32151	0.23320	0.18692	0.25838
gg	0.26746	0.29268	0.22352	0.21634
gt	0.21600	0.18334	0.26213	0.33853
ta	0.35342	0.16646	0.15262	0.32750
tc	0.29702	0.21931	0.22825	0.25543
tg	0.23549	0.26070	0.23044	0.27338
tt	0.22564	0.18481	0.21803	0.37151

Markov models show strong variations between organisms

Saccharomyces cerevisiae (Fungus)

Pr	a	c	g	t
a	0.37000	0.16588	0.17908	0.28504
c	0.32610	0.19058	0.16818	0.31514
g	0.31163	0.21456	0.18957	0.28424
t	0.27256	0.17991	0.17364	0.37389

Escherichia coli K12 (Proteobacteria)

pr	a	c	g	t
a	0.34491	0.18156	0.17676	0.29677
c	0.30806	0.21557	0.22129	0.25507
g	0.27123	0.25972	0.21545	0.25360
t	0.24080	0.19176	0.21144	0.35599

Mycobacterium leprae (Actinobacteria)

pr	a	c	g	t
a	0.23239	0.28694	0.25692	0.22375
c	0.24574	0.24601	0.30574	0.20252
g	0.21748	0.29238	0.25535	0.23479
t	0.18806	0.26081	0.31784	0.23329

Mycoplasma genitalium (Firmicute, intracellular)

pr	a	c	g	t
a	0.45565	0.11743	0.13602	0.29091
c	0.39457	0.13008	0.06403	0.41132
g	0.31505	0.18738	0.12047	0.37710
t	0.32450	0.09573	0.11934	0.46044

Bacillus subtilis (Firmicute, extracellular)

pr	a	c	g	t
a	0.38159	0.13935	0.18767	0.29139
c	0.33699	0.19499	0.16508	0.30293
g	0.34249	0.18100	0.23541	0.24110
t	0.25122	0.17199	0.19402	0.38278

Plasmodium falciparum (Aplicomplexa, intracellular)

pr	a	c	g	t
a	0.39821	0.06446	0.05206	0.48527
c	0.47798	0.13336	0.06695	0.32171
g	0.36764	0.08587	0.12431	0.42217
t	0.44739	0.05676	0.07673	0.41912

Anopheles gambiae (Insect)

pr	a	c	g	t
a	0.34603	0.21388	0.18890	0.25119
c	0.31499	0.21232	0.24159	0.23109
g	0.26036	0.25414	0.20275	0.28275
t	0.20368	0.20710	0.24970	0.33951

Homo sapiens (Mammalian)

pr	a	c	g	t
a	0.29760	0.19031	0.28856	0.22353
c	0.28019	0.30209	0.11692	0.30080
g	0.24408	0.24738	0.30309	0.20545
t	0.18589	0.23061	0.27491	0.30859

Scoring a sequence segment with a Markov model

- The example below illustrates the computation of the probability of a seqence segment (CCTACTATATGCCCAGAATT) with a Markov chain of order 2, calibrated from 3nt frequencies on the yeast genome.

$$P(S) = P(S_{1,m}) \prod_{i=m+1}^{L} P(r_i \mid S_{i-m,i-1})$$

Transition matrix, order 2

Prefix/Suffix	A	C	G	T	P(Prefix)	N(Prefix)
AA	0.388	0.161	0.200	0.251	**0.112**	**525,000**
AC	0.339	0.198	0.173	0.290	**0.054**	**251,072**
AG	0.345	0.204	0.196	0.255	**0.059**	**274,601**
AT	0.311	0.184	0.182	0.323	**0.088**	**413,946**
CA	0.347	0.178	0.189	0.286	**0.063**	**293,750**
CC	0.341	0.190	0.161	0.309	**0.038**	**178,110**
CG	0.293	0.221	0.196	0.290	**0.031**	**145,876**
CT	0.229	0.195	0.205	0.371	**0.059**	**275,634**
GA	0.394	0.155	0.187	0.264	**0.059**	**277,053**
GC	0.330	0.205	0.169	0.297	**0.039**	**184,192**
GG	0.318	0.217	0.187	0.277	**0.037**	**173,266**
GT	0.285	0.175	0.204	0.336	**0.051**	**239,384**
TA	0.300	0.193	0.168	0.339	**0.079**	**369,426**
TC	0.313	0.203	0.152	0.332	**0.060**	**280,131**
TG	0.302	0.209	0.208	0.282	**0.060**	**279,783**
TT	0.210	0.208	0.189	0.392	**0.111**	**520,906**
P(Suffix)	**0.313**	**0.191**	**0.187**	**0.310**		
N(suffix)	**1,466,075**	**893,444**	**873,260**	**1,449,351**		

pos	P(R|W)		wR	S	P(S)
1	P(CC)	0.038	cc	CC	3.80E-02
2	P(T|CC)	0.309	ccT	CCT	1.17E-02
3	P(A|CT)	0.229	ctA	CCTA	2.69E-03
4	P(C|TA)	0.193	taC	CCTAC	5.19E-04
5	P(T|AC)	0.290	acT	CCTACT	1.50E-04
6	P(A|CT)	0.229	ctA	CCTACTA	3.45E-05
7	P(T|TA)	0.339	taT	CCTACTAT	1.17E-05
8	P(A|AT)	0.311	atA	CCTACTATA	3.63E-06
9	P(T|TA)	0.339	taT	CCTACTATAT	1.23E-06
10	P(G|AT)	0.182	atG	CCTACTATATG	2.25E-07
11	P(C|TG)	0.209	tgC	CCTACTATATGC	4.69E-08
12	P(C|GC)	0.205	gcC	CCTACTATATGCC	9.61E-09
13	P(C|CC)	0.190	ccC	CCTACTATATGCCC	1.82E-09
14	P(A|CC)	0.341	ccA	CCTACTATATGCCCA	6.21E-10
15	P(G|CA)	0.189	caG	CCTACTATATGCCCAG	1.17E-10
16	P(A|AG)	0.345	agA	CCTACTATATGCCCAGA	4.04E-11
17	P(A|GA)	0.394	gaA	CCTACTATATGCCCAGAA	1.59E-11
18	P(T|AA)	0.251	aaT	CCTACTATATGCCCAGAAT	4.00E-12
19	P(T|AT)	0.323	atT	CCTACTATATGCCCAGAATT	1.29E-12

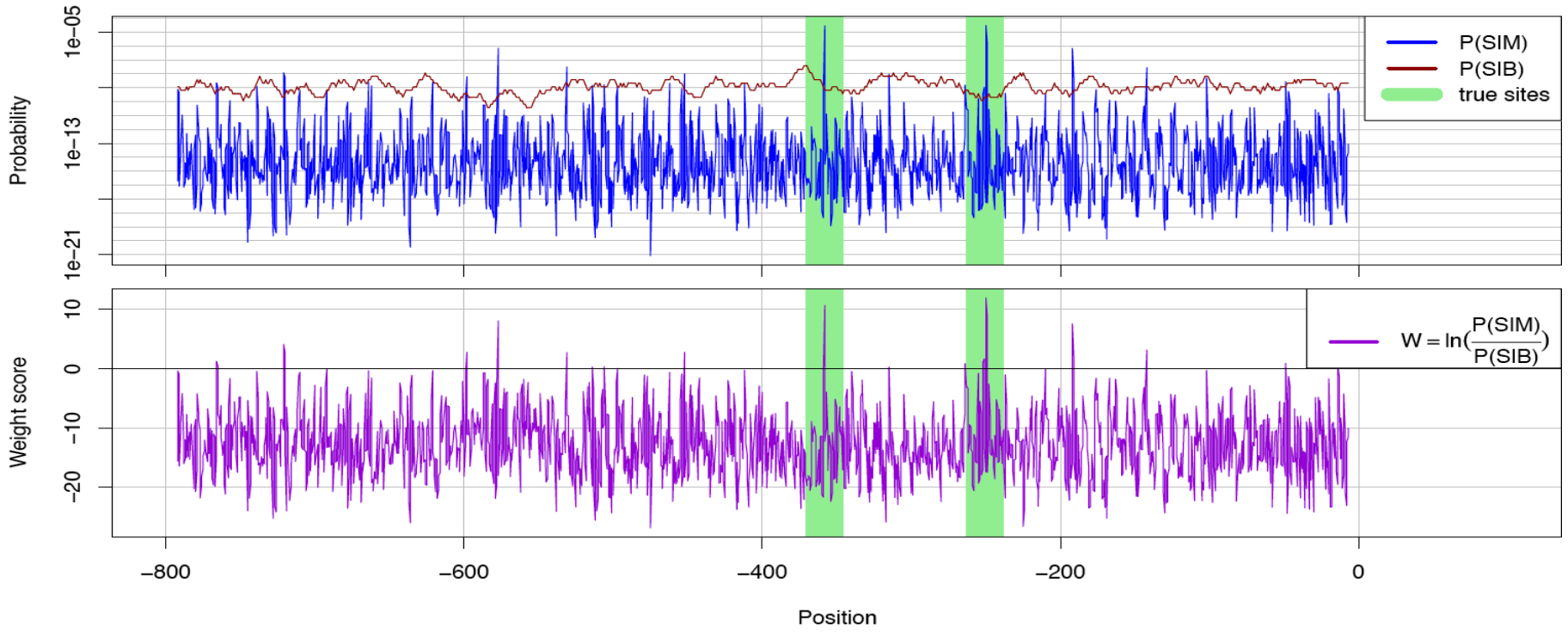
Scanning a sequence with a position-specific scoring matrix

- **P(S|M)** probability for site S to be generated as an instance of the motif.
- **P(S|B)** probability for site S to be generated as an instance of the background.
- **W** weight, i.e. the log ratio of the two above probabilities.
  - A positive weight indicates that a site is more likely to be an instance of the motif than of the background.

$$P(S \mid M) = \prod_{j=1}^{w} f'_{r_j j}$$

$$P(S \mid B) = \prod_{j=1}^{w} p_{r_j}$$

$$W_S = \ln\left(\frac{P(S \mid M)}{P(S \mid B)}\right)$$



Sand, O., Turatsinze, J.V. and van Helden, J. (2008). Evaluating the prediction of cis-acting regulatory elements in genome sequences In Frishman, D. and Valencia, A. (eds.), Modern genome annotation: the BioSapiens network. Springer.

Question

- If we use a position-specific scoring matrix to predict binding sites for a transcription factor, how good is this matrix to recognize *bona fide* binding sites from other sequences (background) ?

- We will have to measure the trade-off between two **qualities** of a matrix
  - **Sensitivity** (**=coverage**): which fraction of the actual binding site are we able to detect ?
  - **Positive predictive power (PPV)**: which fraction of our predictions do correspond to actual binding sites ?
- Several **parameters** have a crucial importance for determining the reliability of the predictions
  - **The matrix**: which is the best matrix that can be built from a collection of annotated binding sites ?
    - Selection of the binding sites to be incorporated
    - Alignment procedure.
    - Matrix width.
    - Background model used for building the matrix
  - Scanning parameters:
    - Background model used for scanning the sequences,
    - Score threshold.

Approaches

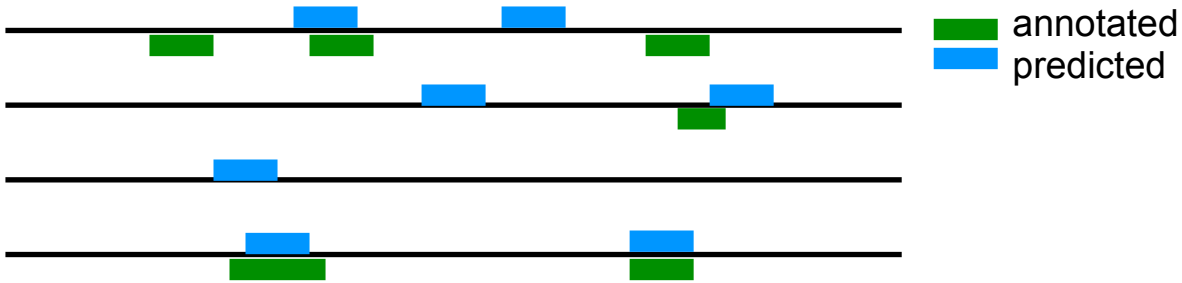
- Theoretical estimates
  - Various statistics have been used to estimate the relevance of a motif
    - Relative entropy (Schneider, 1986)
    - Information Content (Hertz & Stormo, 1990; 1999)
    - P-value (Staden, 1989) or E-value (Hertz & Stormo, 1999; Bailey, 1994)
    - Consensus score (Thijs et al., 2001)
    - Log-likelihood (Thijs et al., 2001)
    - Alpha- and beta-riks (Rahmann et al, 2003)
  - All those statistics rely on theoretical assumption, which might fail to reproduce the characteristics of real biological sequences.
- We propose here a complementary approach, combining theoretical and empirical measurements of matrix quality.
  - Compute the theoretical distributions of scores for all possible threshold values.
  - Compare theoretical distributions with those obtained in several reference data sets.
    - Positive control
      - Annotated binding sites for the transcription factor of interest
      - Cis-regulatory regions of the target genes
    - Negative control
      - DNA sequences supposed to be devoid of binding sites.

Rahmann, S., Muller, T. & Vingron, M. (2003). On the power of profiles for transcription factor binding site detection. Stat Appl Genet Mol Biol 2, Article7. PMID 16646785

Sensitivity, PPV and Accuracy curves

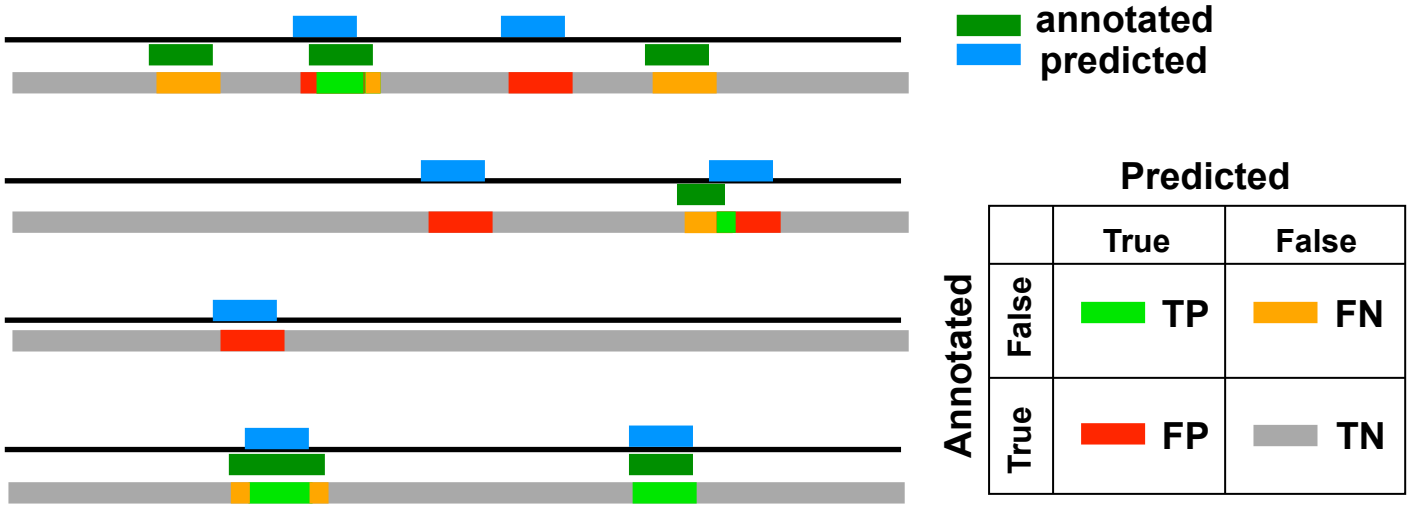
Comparison between annotated and predicted sites

- The annotated and predicted sites are compared.



Comparison at the nucleotide level

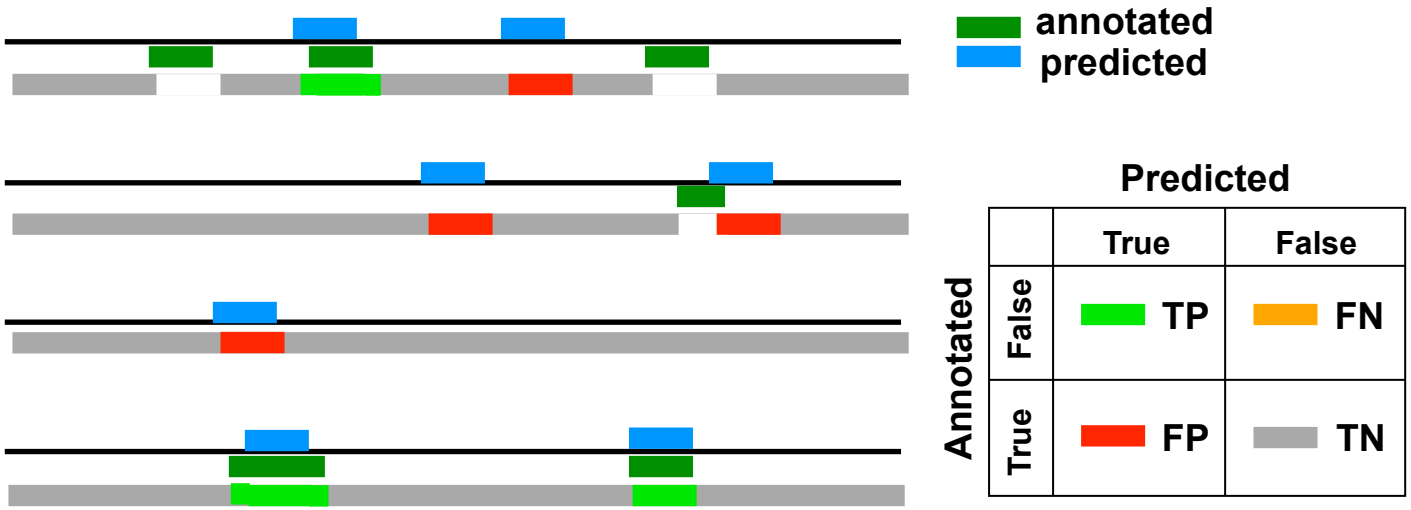
- Annotated and predicted sites can be compared **at the nucleotide level**.
- Each predicted nucleotide is considered as a match if it falls within an annotated site.
- A predicted site can thus contain a mixture of false positive and true positive nucleotides.
- An annotated site can thus contain a mixture of true positive and false negative nuceotides.



TP	True Positive	Annotated and predicted.
FP	False Positive	Predicted but not annotated
TN	True Negative	Neither annotated nor predicted
FN	False Negative	Annotated but not predicted

Comparison at the site level

- Annotated and predicted sites can be compared **at the site level**.
- Each predicted site is considered as a match (**as a whole**) if it overlaps with an annotated site.
- Each annotated site is considered as detected (**as a whole**) if it is matched by at least one predicted site.
- A threshold can be imposed on the minimal number of overlapping nucleotides in order to consider that a predicted site does or not match an annotated site.

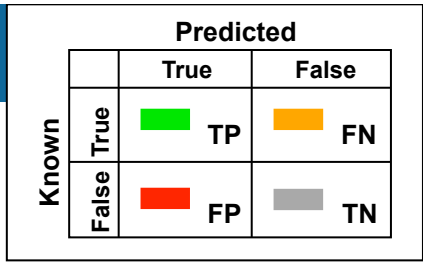


TP	True Positive	Annotated and predicted.
FP	False Positive	Predicted but not annotated
TN	True Negative	Neither annotated nor predicted
FN	False Negative	Annotated but not predicted

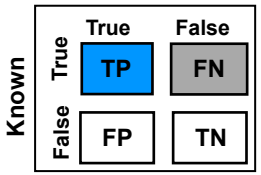
Validation statistics

- Various statistics can be derived from the 4 elements of a contingency table.

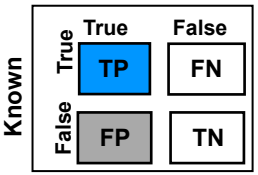
Abbrev	Name	Formula
TP	True positive	TP
FP	False positive	FP
FN	False negative	FN
TN	True negative	TN
KP	Known Positive	TP+FN
KN	Known Negative	TN+FP
PP	Predicted Positive	TP+FP
PN	Predicted Negative	FN+TN
N	Total	TP + FP + FN + TN
Prev	Prevalence	(TP + FN)/N
ODP	Overall Diagnostic Power	(FP + TN)/N
CCR	Correct Classification Rate	(TP + TN)/N
Sn	**Sensitivity**	**TP/(TP + FN)**
Sp	Specificity	TN/(FP + TN)
FPR	False Positive Rate	FP/(FP + TN)
FNR	False Negative Rate	FN/(TP + FN)
PPV	**Positive Predictive Value**	**TP/(TP + FP)**
FDR	False Discovery Rate	FP/(FP+TP)
NPV	Negative Predictive Value	TN/(FN + TN)
Mis	Misclassification Rate	(FP + FN)/N
Odds	Odds-ratio	(TP + TN)/(FN + FP)
Kappa	Kappa	((TP + TN) - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))/(N - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))
NMI	NMI n(s)	(1 - -TP*log(TP)-FP*log(FP)-FN*log(FN)-TN*log(TN)+(TP+FP)*log(TP+FP)+(FN+TN)*log(FN+TN))/(N*log(N) - ((TP+FN)*log(TP+FN) + (FP+TN)*log(FP+TN)))
ACP	Average Conditional Probability	0.25*(Sn+ PPV + Sp + NPV)
MCC	Matthews correlation coefficient	(TP*TN - FP*FN) / sqrt[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)]
Acc.a	Arithmetic accuracy	(Sn + PPV)/2
Acc.a2	Accuracy (alternative)	(Sn + Sp)/2
Acc.g	Accuracy (Sn+PPV)/2	sqrt(Sn*PPV)
Hit.noTN	A sort of hit rate without TN (to avoid the effect of their large number)	TP/(TP+FP+FN)



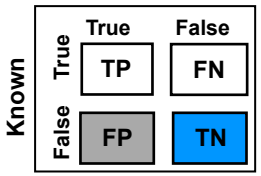
Sn = TP/(TP+FN)



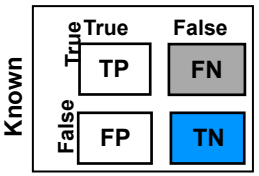
PPV=TP/(TP+FP)



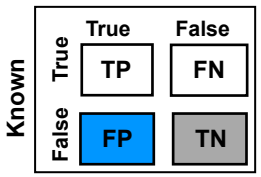
Sp=TN/(FP+TN)



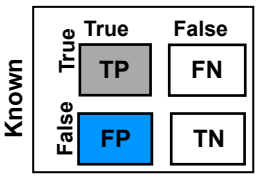
NPV=TN/(FN+TN)



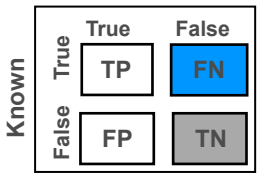
FPR=FP/(FP+TN)



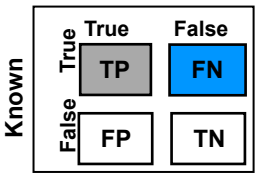
FDR=FP/(FP+TP)



FN/(FN+TN)



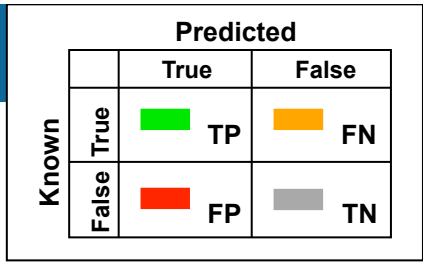
FNR=FN/(TP+FN)



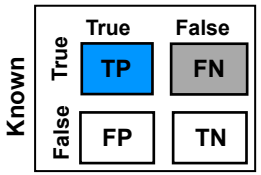
Validation statistics including TN should be avoided

- For the predictions of TF binding sites, TN can represent >99.9%
- -> All the statistics including TN are misleading (e.g. Sp can be very high)

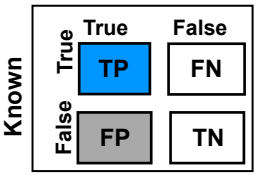
Abbrev	Name	Formula
TP	True positive	TP
FP	False positive	FP
FN	False negative	FN
TN	True negative	TN
KP	Known Positive	TP+FN
KN	Known Negative	TN+FP
PP	Predicted Positive	TP+FP
PN	Predicted Negative	FN+TN
N	Total	TP + FP + FN + TN
Prev	Prevalence	(TP + FN)/N
ODP	Overall Diagnostic Power	(FP + TN)/N
CCR	Correct Classification Rate	(TP + TN)/N
Sn	**Sensitivity**	**TP/(TP + FN)**
Sp	Specificity	TN/(FP + TN)
FPR	False Positive Rate	FP/(FP + TN)
FNR	False Negative Rate	FN/(TP + FN)
PPV	**Positive Predictive Value**	**TP/(TP + FP)**
FDR	False Discovery Rate	FP/(FP+TP)
NPV	Negative Predictive Value	TN/(FN + TN)
Mis	Misclassification Rate	(FP + FN)/N
Odds	Odds-ratio	(TP + TN)/(FN + FP)
Kappa	Kappa	((TP + TN) - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))/(N - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))
NMI	NMI n(s)	(1 - -TP*log(TP)-FP*log(FP)-FN*log(FN)-TN*log(TN)+(TP+FP)*log(TP+FP)+(FN+TN)*log(FN+TN))/(N*log(N) - ((TP+FN)*log(TP+FN) + (FP+TN)*log(FP+TN)))
ACP	Average Conditional Probability	0.25*(Sn+ PPV + Sp + NPV)
MCC	Matthews correlation coefficient	(TP*TN - FP*FN) / sqrt[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)]
Acc.a	Arithmetic accuracy	(Sn + PPV)/2
Acc.a2	Accuracy (alternative)	(Sn + Sp)/2
Acc.g	Accuracy (Sn+PPV)/2	sqrt(Sn*PPV)
Hit.noTN	A sort of hit rate without TN (to avoid the effect of their large number)	TP/(TP+FP+FN)



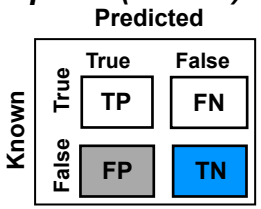
Sn = TP/(TP+FN)



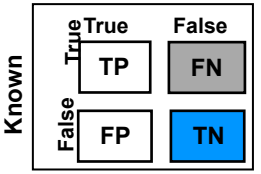
PPV=TP/(TP+FP)



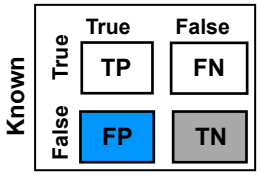
Sp=TN/(FP+TN)



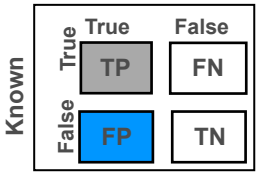
NPV=TN/(FN+TN)



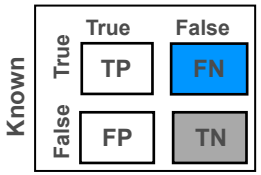
FPR=FP/(FP+TN)



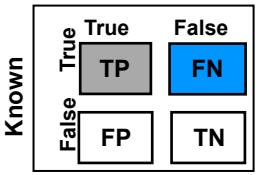
FDR=FP/(FP+TP)



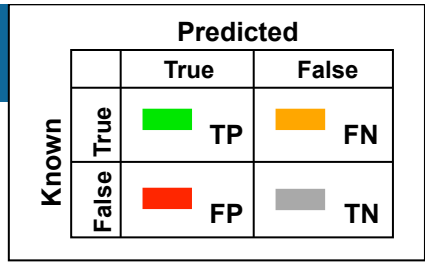
FN/(FN+TN)



FNR=FN/(TP+FN)

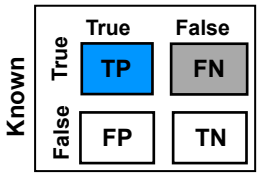


The arithmetic accuracy should be avoided

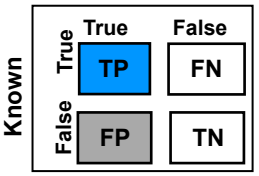


Abbrev	Name	Formula
TP	True positive	TP
FP	False positive	FP
FN	False negative	FN
TN	True negative	TN
KP	Known Positive	TP+FN
KN	Known Negative	TN+FP
PP	Predicted Positive	TP+FP
PN	Predicted Negative	FN+TN
N	Total	TP + FP + FN + TN
Prev	Prevalence	(TP + FN)/N
ODP	Overall Diagnostic Power	(FP + TN)/N
CCR	Correct Classification Rate	(TP + TN)/N
Sn	**Sensitivity**	**TP/(TP + FN)**
Sp	Specificity	TN/(FP + TN)
FPR	False Positive Rate	FP/(FP + TN)
FNR	False Negative Rate	FN/(TP + FN)
PPV	**Positive Predictive Value**	**TP/(TP + FP)**
NPV	Negative Predictive Value	TN/(FN + TN)
Mis	Misclassification Rate	(FP + FN)/N
Odds	Odds-ratio	(TP + TN)/(FN + FP)
Kappa	Kappa	((TP + TN) - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))/(N - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))
NMI	NMI n(s)	(1 - -TP*log(TP)-FP*log(FP)-FN*log(FN)-TN*log(TN)+(TP+FP)*log(TP+FP)+(FN+TN)*log(FN+TN))/(N*log(N) - ((TP+FN)*log(TP+FN) + (FP+TN)*log(FP+TN)))
ACP	Average Conditional Probability	0.25*(Sn+ PPV + Sp + NPV)
MCC	Matthews correlation coefficient	(TP*TN - FP*FN) / sqrt[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)]
Acc.a	Arithmetic accuracy	(Sn + PPV)/2
Acc.a2	Accuracy (alternative)	(Sn + Sp)/2
Acc.g	**Accuracy (Sn+PPV)/2**	**sqrt(Sn*PPV)**
Hit.noTN	A sort of hit rate without TN (to avoid the effect of their large number)	TP/(TP+FP+FN)

Sn = TP/(TP+FN)



PPV=TP/(TP+FP)



- It is easy to be fooled by the arithmetic accuracy: let us imagine that a trivial program predicts all features as positive
  - **-> *Sn* guaranteed to be 100%**
  - Of course, you have a poor PPV
  - $Acc_a = (Sn + PPV)/2$
    - -> guaranteed to be >50%
- The geometric accuracy circumvents this problem
  - $Acc_g = sqrt(Sn*PPV)$
  - Requires for **both** *Sn* and *PPV* to be high.

Accuracy profiles for the human CREB factor

- We can measure for each score value
  - Sn = TP/(TP+FN)
  - PPV=TP/(TP/+FP)
  - Acc.g =sqrt(Sn*PPV)
- *With increaseing scores*
  - *Sn* decreases when score increases (we are progressively loosing sites)
  - If the matrix + annotations are good, the *PPV* should increase.
  - The accuracy shows some optimum (peak)
- Beware
  - There is a bias in these evaluations.
  - The "optimal" accuracy may be catastrophic choice for whole-genome predictions.
  - Indeed, even an apparently low risk of false positive (FPR=0.1%) represents a high number of false predictions, since it is multiplied by the size of the genome.

Ws=-25

Annotated	Predicted TRUE	Predicted FALSE
TRUE	25	0
FALSE	22515	0

Sn 100.0%
PPV 0.1%
Acc.g 3.3%

Ws=0

Annotated	Predicted TRUE	Predicted FALSE
TRUE	24	1
FALSE	181	22334

Sn 96.0%
PPV 11.7%
Acc.g 33.5%

Ws=4.119

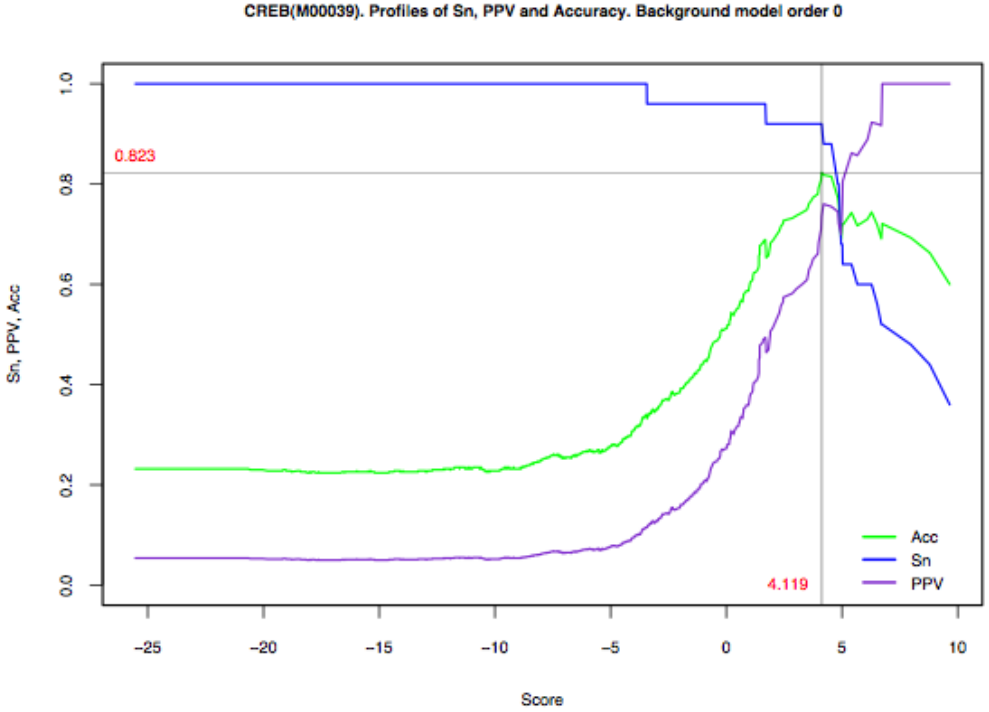
Annotated	Predicted TRUE	Predicted FALSE
TRUE	23	2
FALSE	30	22485

Sn 92.0%
PPV 43.4%
Acc.g 63.2%

Ws=8

Annotated	Predicted TRUE	Predicted FALSE
TRUE	11	14
FALSE	7	22508

Sn 44.0%
PPV 61.1%
Acc.g 51.9%

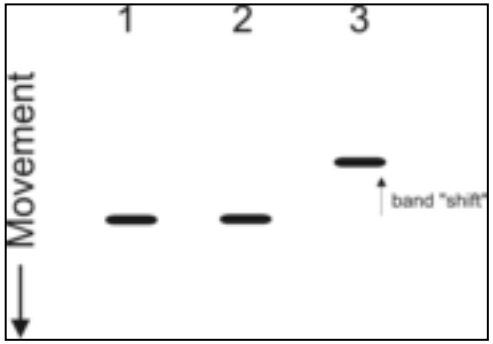


Sand, O., Turatsinze, J. V. and van Helden, J. Evaluating the prediction of cis-acting regulatory elements in genome sequences In Modern genome annotation: the BioSapiens network (Springer, 2008).

Difficulties for estimating site predictions

- Annotations are far from complete
  -> what we consider as false positive can reveal non-annotated binding sites
- The boundaries of binding sites are not defined in an all-or-none fashion.
  - Only a few residues are involved in the direct contact between transcription factor and its binding site, but the flanking residues affect the binding by conformation effects.
- Site boundaries in databases depend much on the experimental evidence
  - Footprint: typically 6-20bp
  - Gel shift: sometimes 50bp or more
- The concept of “binding site” itself can be questioned.
  - Transcription factors have a higher affinity for DNA than for the nucleoplasm.
  - According to some models, they can bind anywhere on DNA, but they spend more time on some sites than on other ones.
  - One could thus consider a continuum of binding affinities.

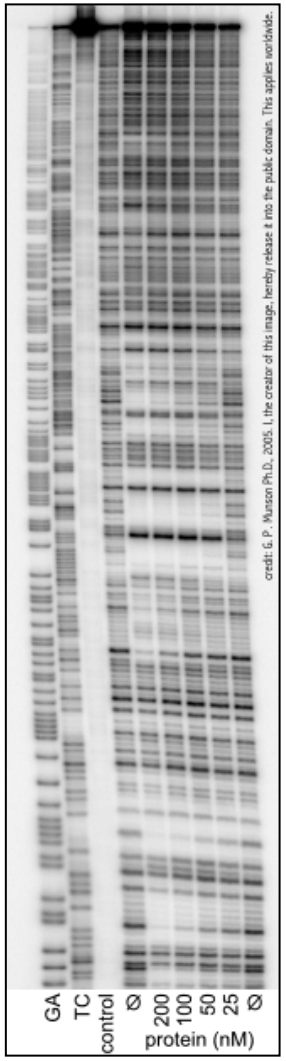
Gel shift assay



Lane 1 is a negative control, and contains only DNA. Lane 2 contains protein as well as a DNA fragment that, based on its sequence, does not interact. Lane 3 contains protein and a DNA fragment that does react; the resulting complex is larger, heavier, and slower-moving. The pattern shown in lane 3 is the one that would result if all the DNA were bound and no dissociation of complex occurred during electrophoresis. When these conditions are not met a second band might be seen in lane 3 reflecting the presence of free DNA or the dissociation of the DNA-protein complex.

Source: Wikipedia

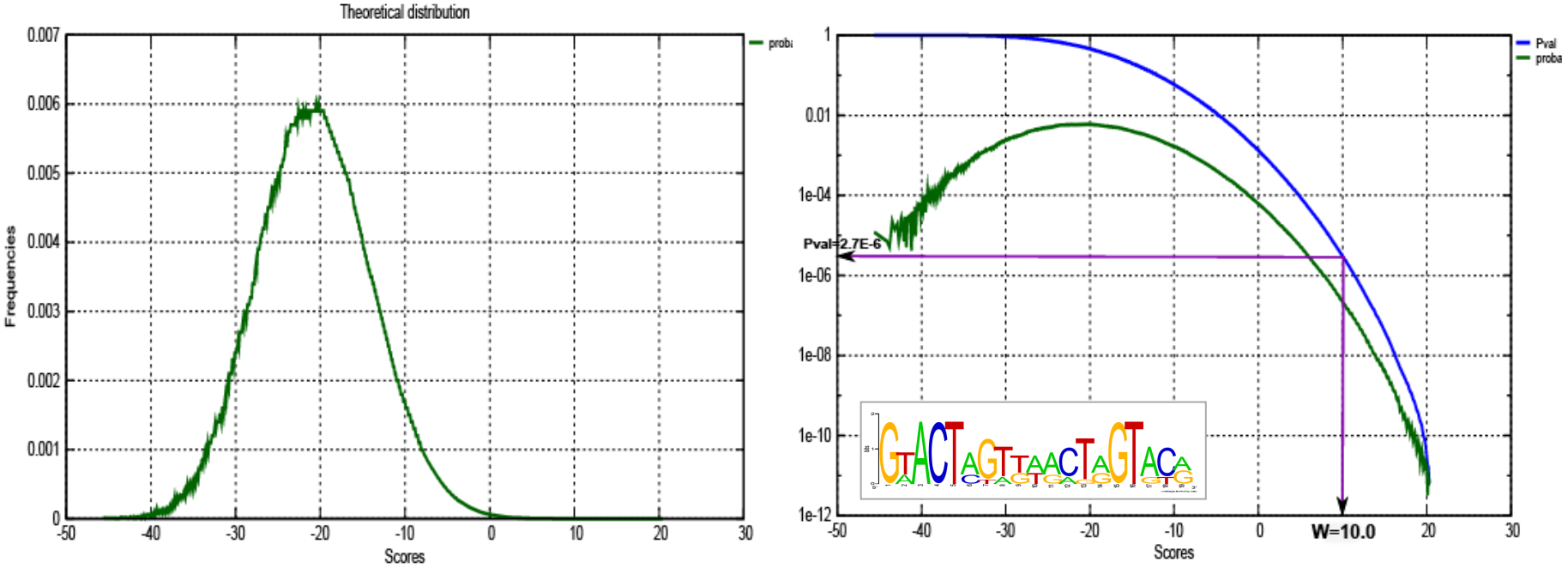
DNAse footprint



Distributions of scores in biological sequences

Theoretical distribution of scores for the TrpR matrix

- The RSAT program *matrix-distrib* allows to compute the theoretical distribution of the weight score according to Markovian background models (by extension of the algorithm from Staden, 1989).
- Green curve
  - Probability $P(W=w)$ to observe ***exactly*** a given weight score at a given position of a random sequence.
- Blue curve
  - Probability $P(W>=w)$ to observe ***at least*** a given weight score at a given position of a random sequence.
  - Note: the Y axis of the right figure is logarithmic.
  - This is an estimate of the ***risk of false positive (FPR)***: probability to consider a position of the sequence as a binding site whereas it is not.
  - Ex: for the TrpR matrix, the probability to observe a score W>=10.0 at a given position is 2.7E-6.



The negative control for the FPR

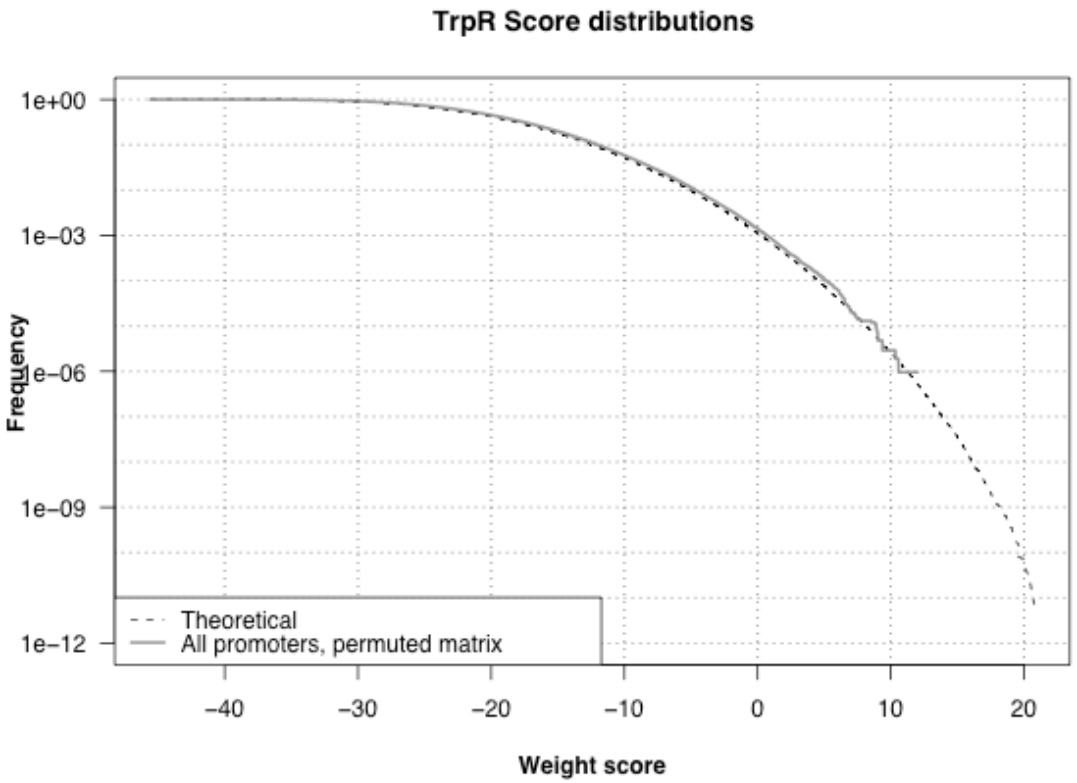
- How reliable is this theoretical distribution ?
  Does it provide a good estimate of the risk of false positives in real conditions ?

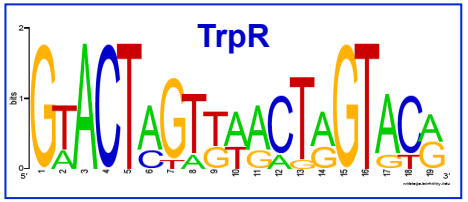
- Approach: scan sequences supposedly devoid of binding sites (“negative” set), and check if the empirical distribution of scores fits the theoretical one.

- Problem: how can we obtain a collection of sequences without binding sites ?
  - Perform experiments to check the non-binding of several thousands/millions of sites ?
    - Problem: costly
  - Generate random sequences ?
    - Problem: this test is trivial, because those random sequences will by definition reflect our theoretical background, and thus fit the theoretical curve.
    - Nothing guarantees that the background model is appropriate to model real biological sequences.
  - Select random sets of genome sequences ?
    - Problem: the selection might include some actual binding sites (bad luck).
  - Scan genome sequences with shuffled matrices.
    - We can shuffle the columns of the matrix, so that the motif becomes uninformative.

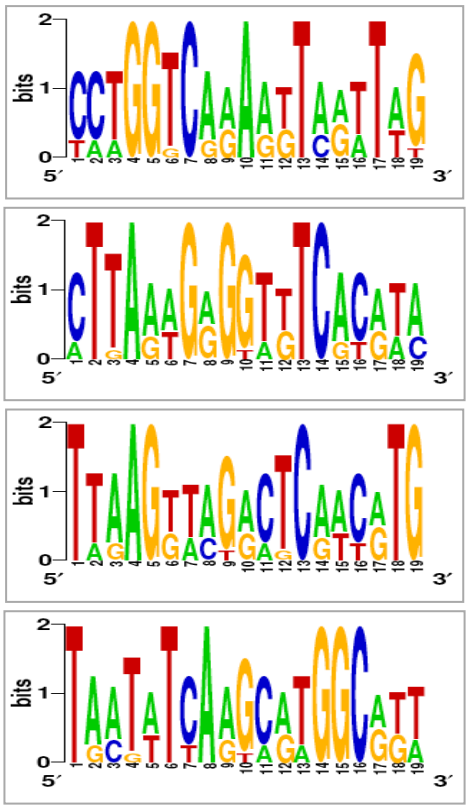
Distribution of matrix scores in E.coli promoters
TrpR matrix from RegulonDB - permuted

- To control the validity of the theoretical distribution: scan a reference sequence set (e.g. all promoters) with shuffled motifs.
  - For this, we perform a random permutation of the columns of the original PSSM, using the program *convert-matrix* (RSAT).
  - This permuted matrix is not supposed to correspond to any existing transcription factor, and we thus expect to find no specific high-scoring site.
  - Indeed, the score distribution observed with the permuted matrix fits the theoretical distribution.



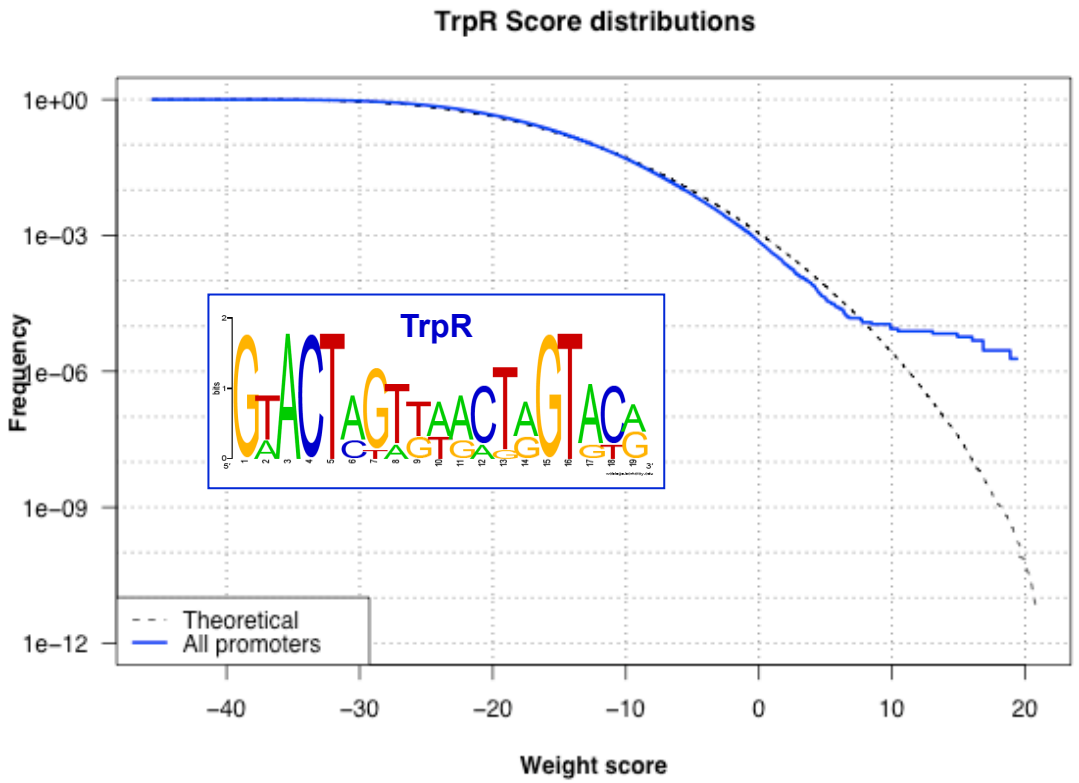


TrpR permutations



...

Distribution of matrix scores in E.coli promoters
TrpR matrix from RegulonDB



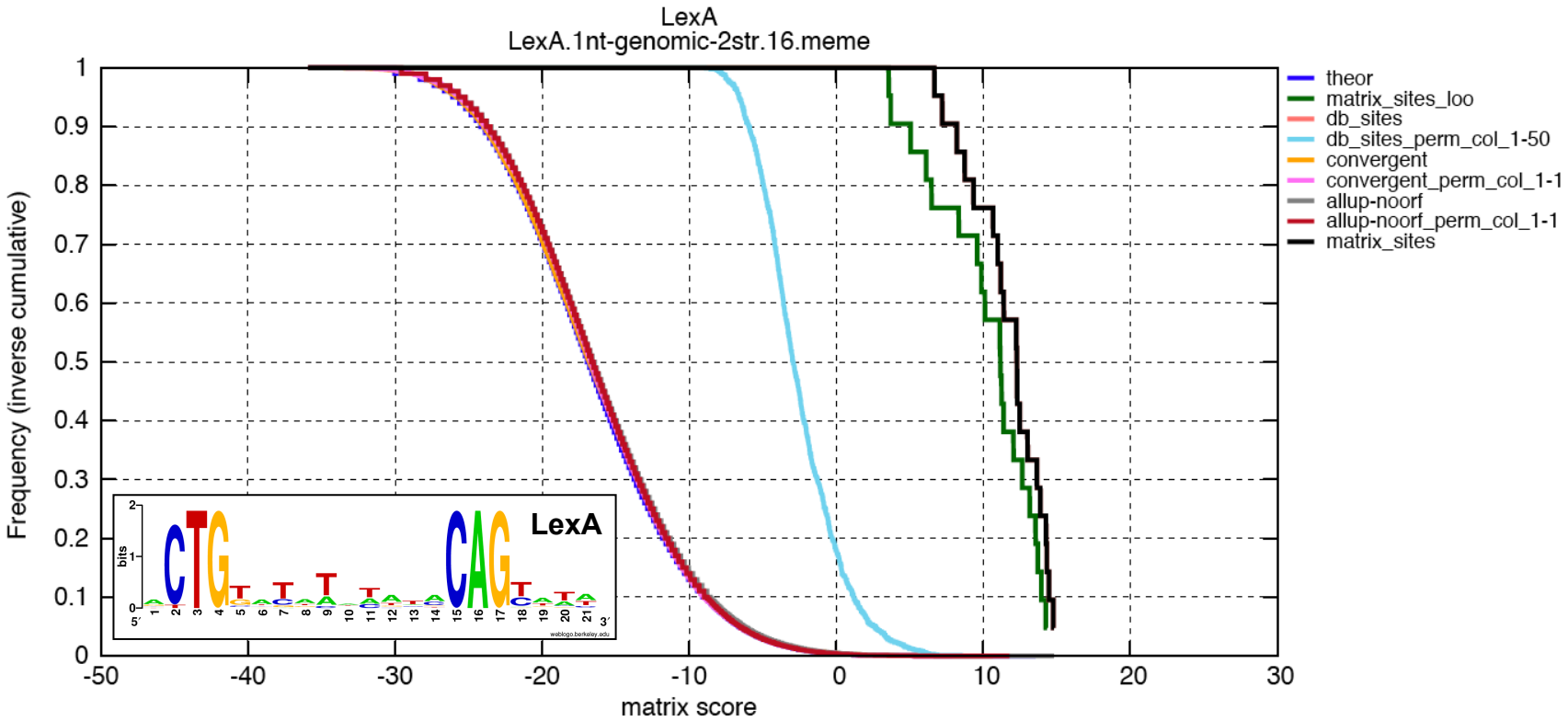
- We compare the distribution observed in all the *E.coli* promoters (blue) with the theoretical distribution, computed by matrix-distrib (blue).
  - For low score values (W <= 8) observed distribution follows pretty well the theoretical one
    - (the small difference is due to a particular of this motif, whcih contains the GTAC tetramer).
  - We also observe a “plateau” corresponding to a few sites having a very high score.
    These high-scoring sites are much more frequent than expected by chance (remind, the axis is logarithmic).

Do these high-scoring sites correspond to the actual binding sites ?

- We can partly check this with the collection of annotated binding sites.

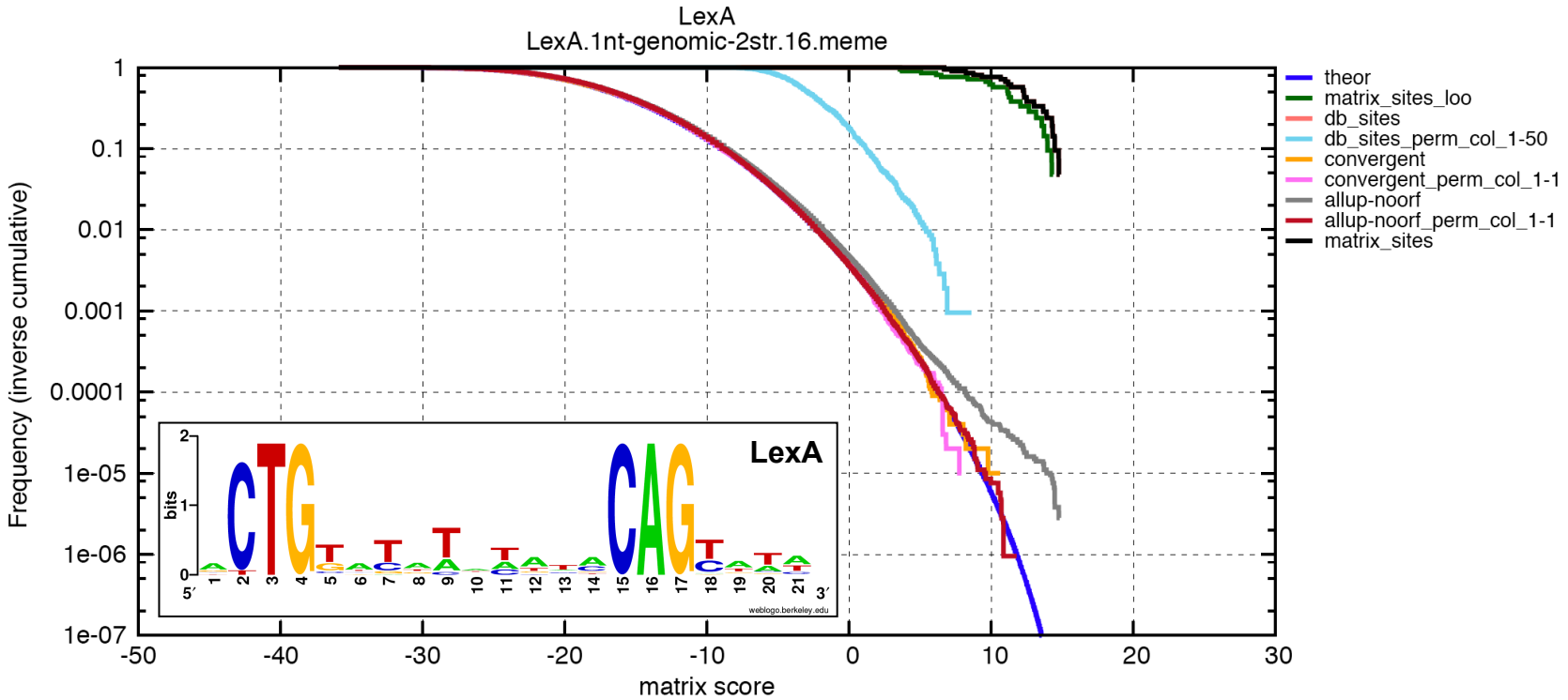
Escherichia coli LexA - Score distributions (logarithmic Y axis)

- \+ **Black**: annotated binding sites that served to build the matrix **(this is biased !)**
- \+ **Green**: leave-one-out (LOO) test with annotated binding sites
- \- **Blue** : theoretical distribution
- \- **Cyan**: scores obtained with 50 permuted matrices in the annotated sites
- \- **Brownish red**: scores obtained with 5 permuted matrices in a random selection of promoters
- -/+ **Gray**: distribution in all promoters

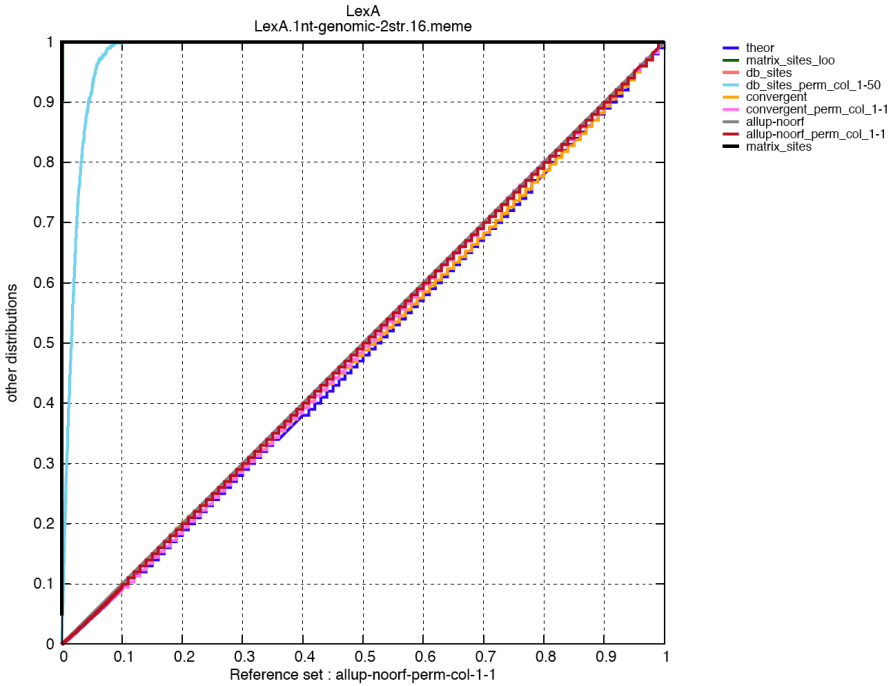


Escherichia coli LexA - Score distributions (logarithmic Y axis)

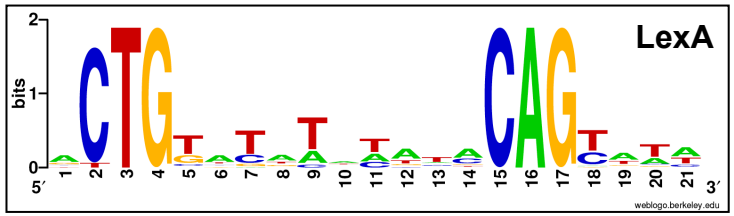
- \+ **Black**: annotated binding sites that served to build the matrix **(this is biased !)**
- \+ **Green**: leave-one-out (LOO) test with annotated binding sites
- \- **Blue** : theoretical distribution
- \- **Cyan**: scores obtained with 50 permuted matrices in the annotated sites
- \- **Brownish red**: scores obtained with 5 permuted matrices in a random selection of promoters
- -/+ **Gray**: distribution in all promoters



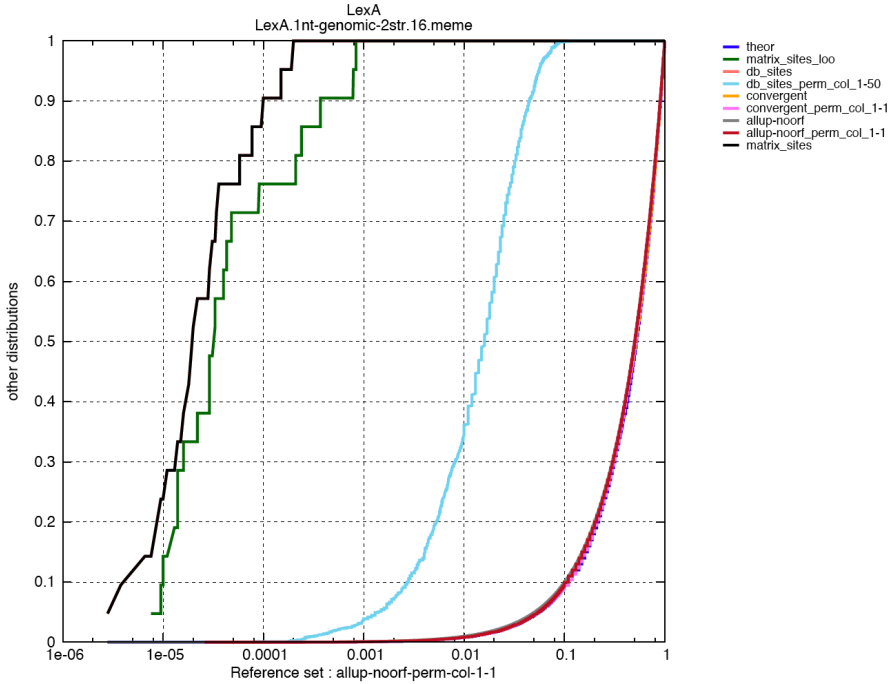
ROC-like curves can be misleading !



- ROC-like representation
- X axis
  - reference negative set
  - = all promoters scanned with permuted matrices
- Y axis: each other set tested
- The curves seem perfect
  - Negative sets are on the diagonal
    - Only exception: permuted sites.
  - This Positive sets are in the top left corner
- However, this is misleading.
- Why ?
  - Because the rate of false positive has to be multiplied by genome scale -> a very small percentage can represent thousands of hits.

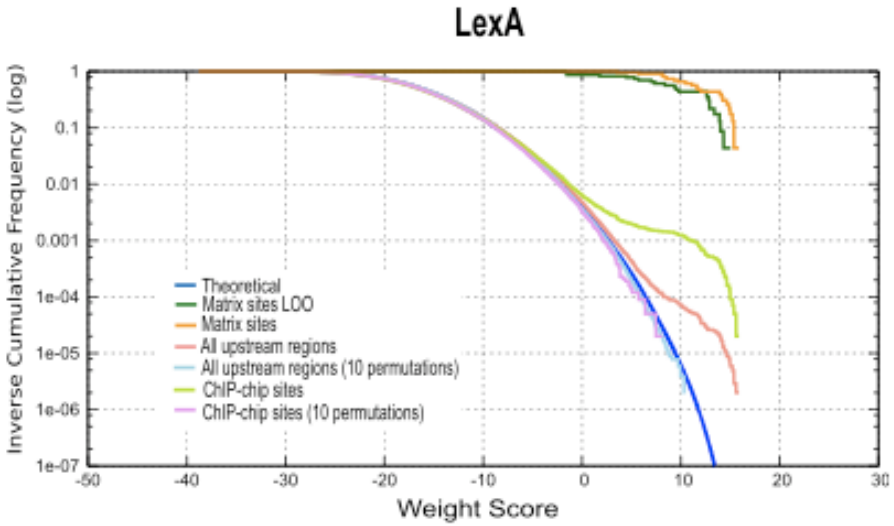


Log-representation of the FPR is more relevant



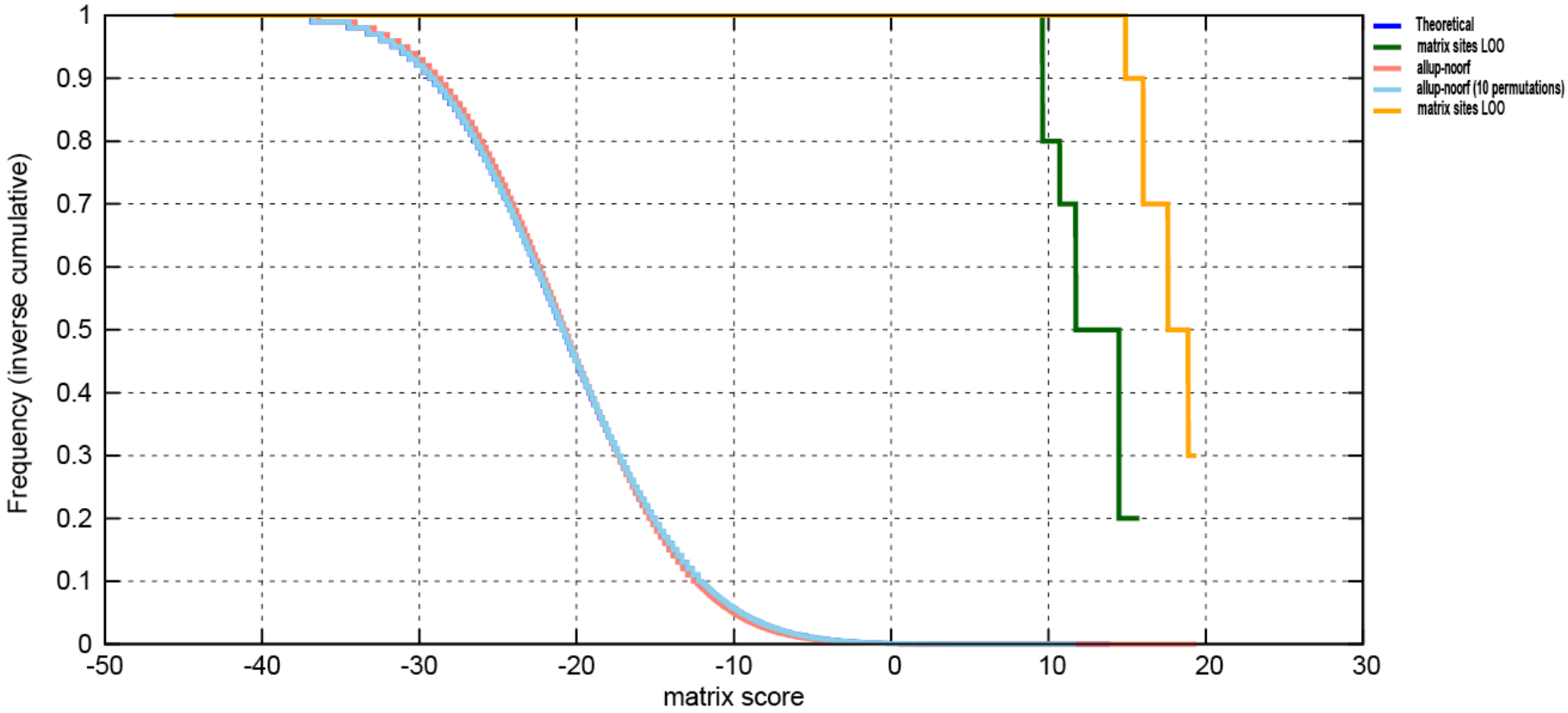
- ROC-like representation
- Logarithmic X axis emphasizes small probabilities
- For a sensitivity of 75%, the FPR is 10^{-4} (green curve)
- This seems quite good, but at the scale of *E.coli g*enome, it represents 400 false predictions !
  - E(FP) = FPR*N
    $=10^{-4}\ 4e^{+6}=400$
- For the human genome, the same FPR would represent
  - $10^{-4}\ 3e^{+9}=300{,}000$ FP !

Comparing weight score distributions

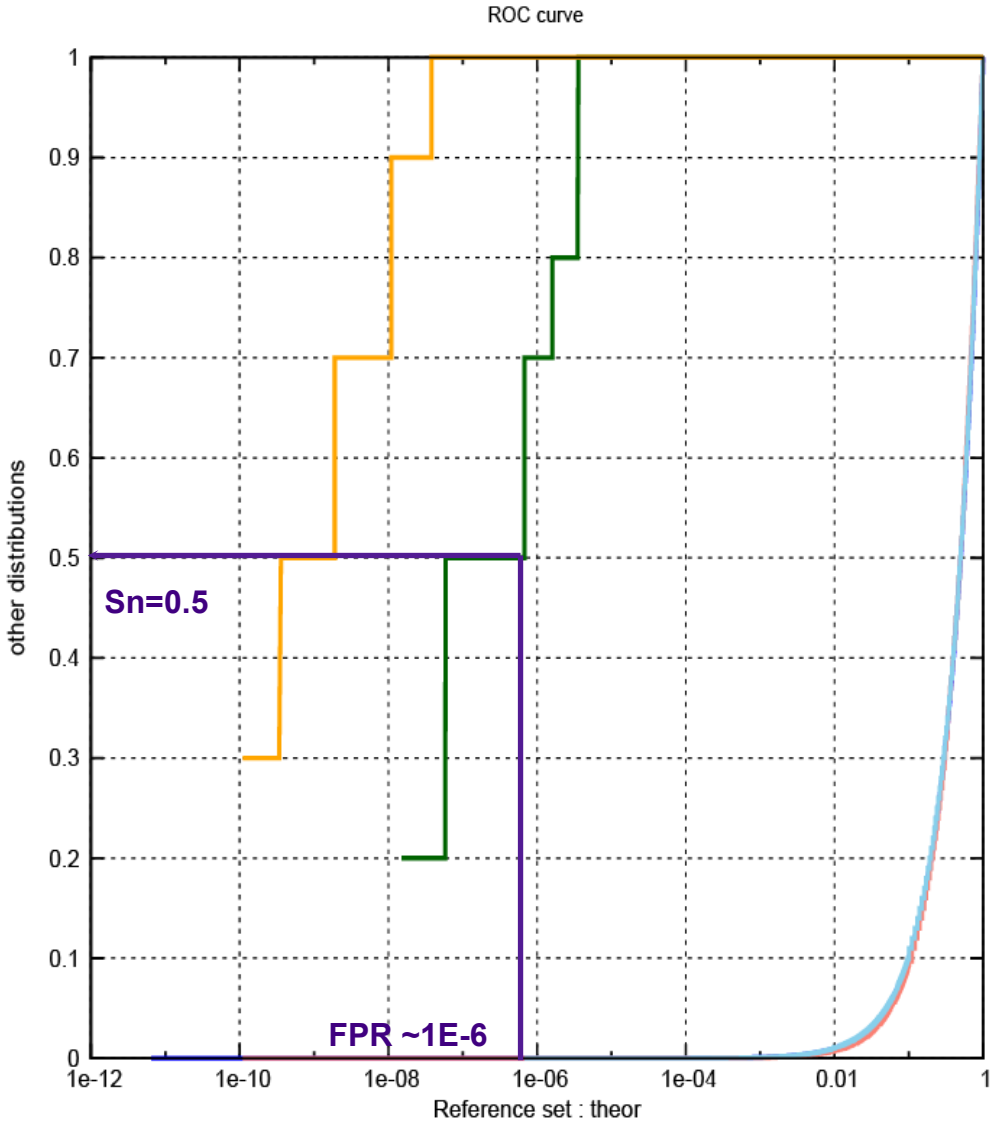


Escherichia coli TrpR - Score distributions in annotated sites

- We compute score distributions in annotated sites
  - with the **full matrix** (BIASED)
  - with **LOO-corrected matrices**

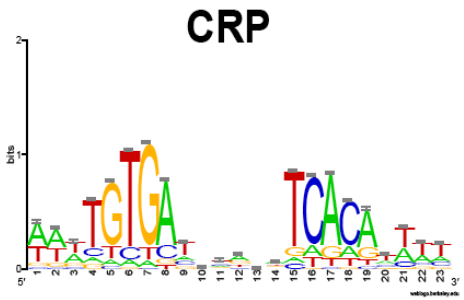


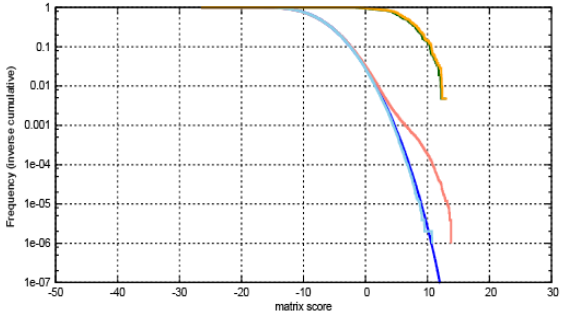
Escherichia coli TrpR - ROC curve

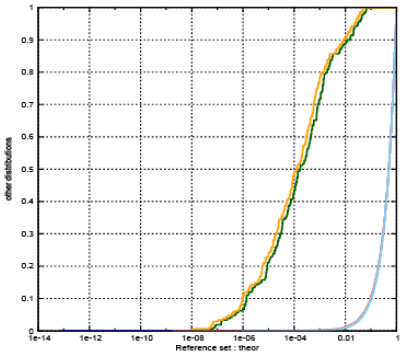


- This representation is derived from the ROC curve, but we display the X axis on a logarithmic scale, in order to emphasize the small ranges of FPR.
- This is particularly important, because for genome-scale searches, the FPR is multiplied by the size of the genome !

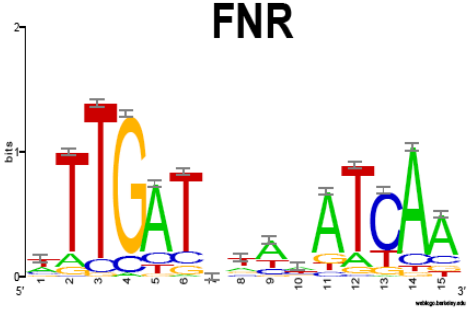
Generic versus specific factors

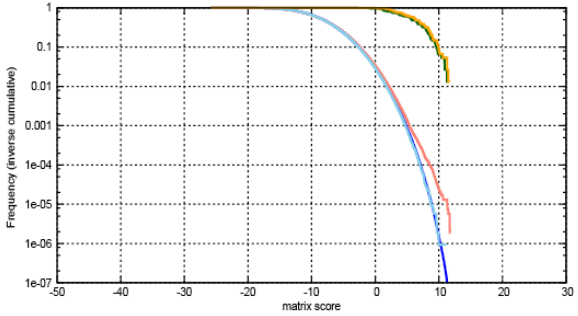


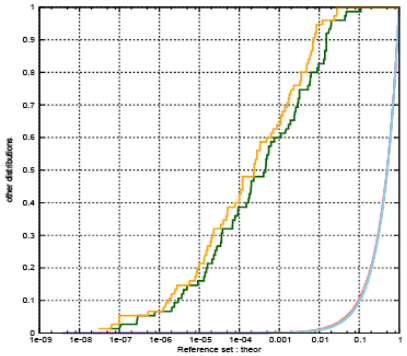




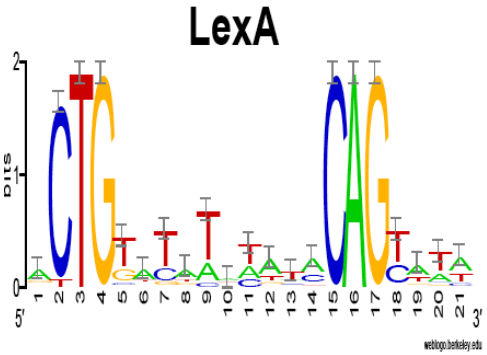
CRP matrix was built from ~200 sites.
The distribution in all promoters discards very early from the theoretical one.
The LOO and biased site distributions are almost identical.

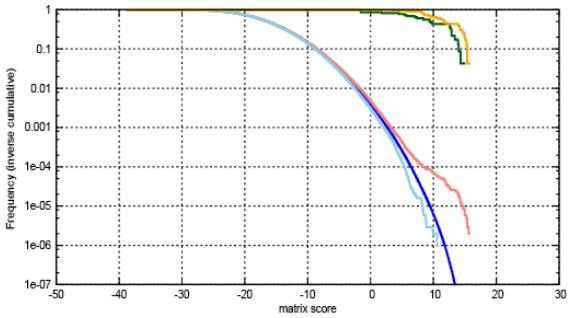


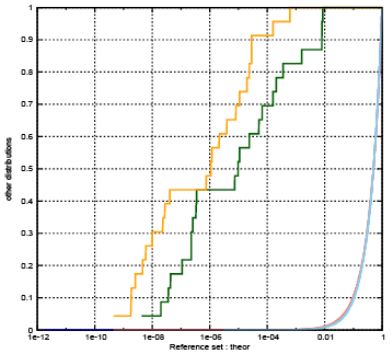




More or less the same effect is observed for FNR.

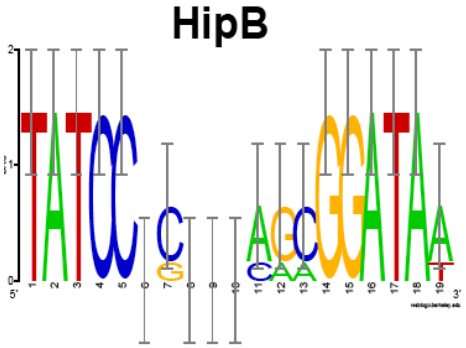


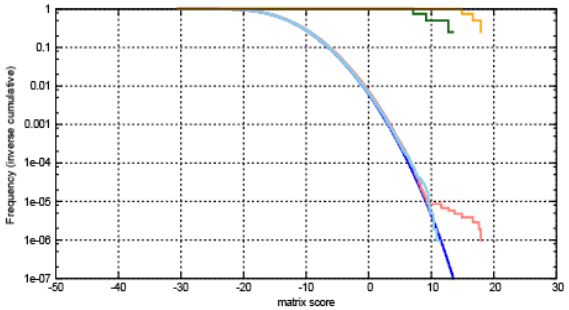


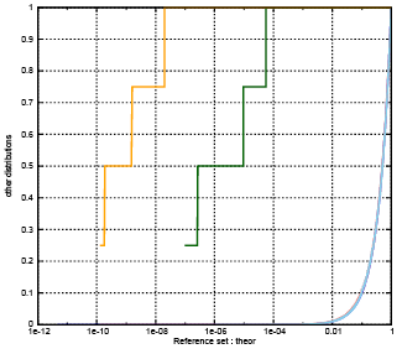


LexA is quite specific, but collecting all sites would represent a high cost:
$FPR_{0.5}$=1.3E^{-5}
$FPR_{0.9}$=8.3E^{-3}

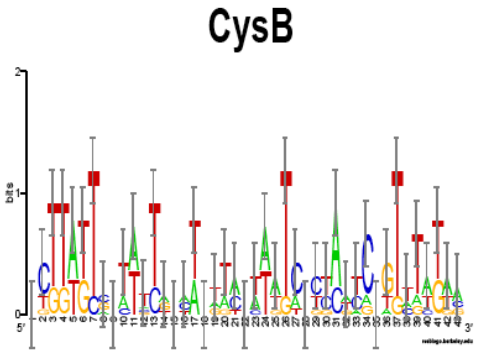
Problematic matrices

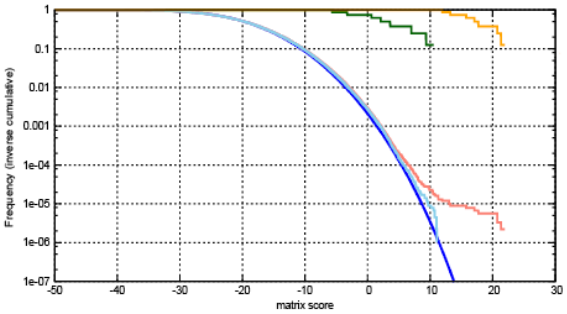


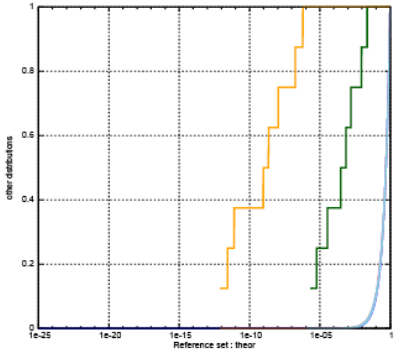




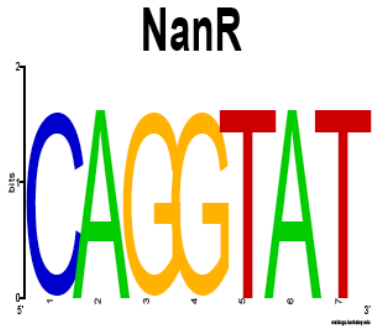
The HipB matrix was built from 3 sites only -> over-fitting.
The biased and LOO FPR curves differ by a factor 10.000

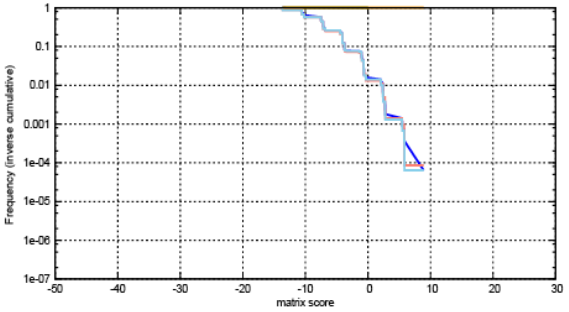


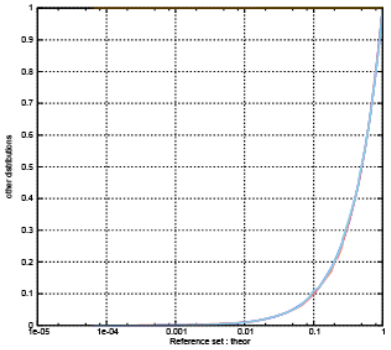




The CysB matrix was built from very large sites (42 columns).
The sites are very far away from the theoretical min Pval (1e-25).
Biased and LOO FPR curves differ by a factor of 100.000.

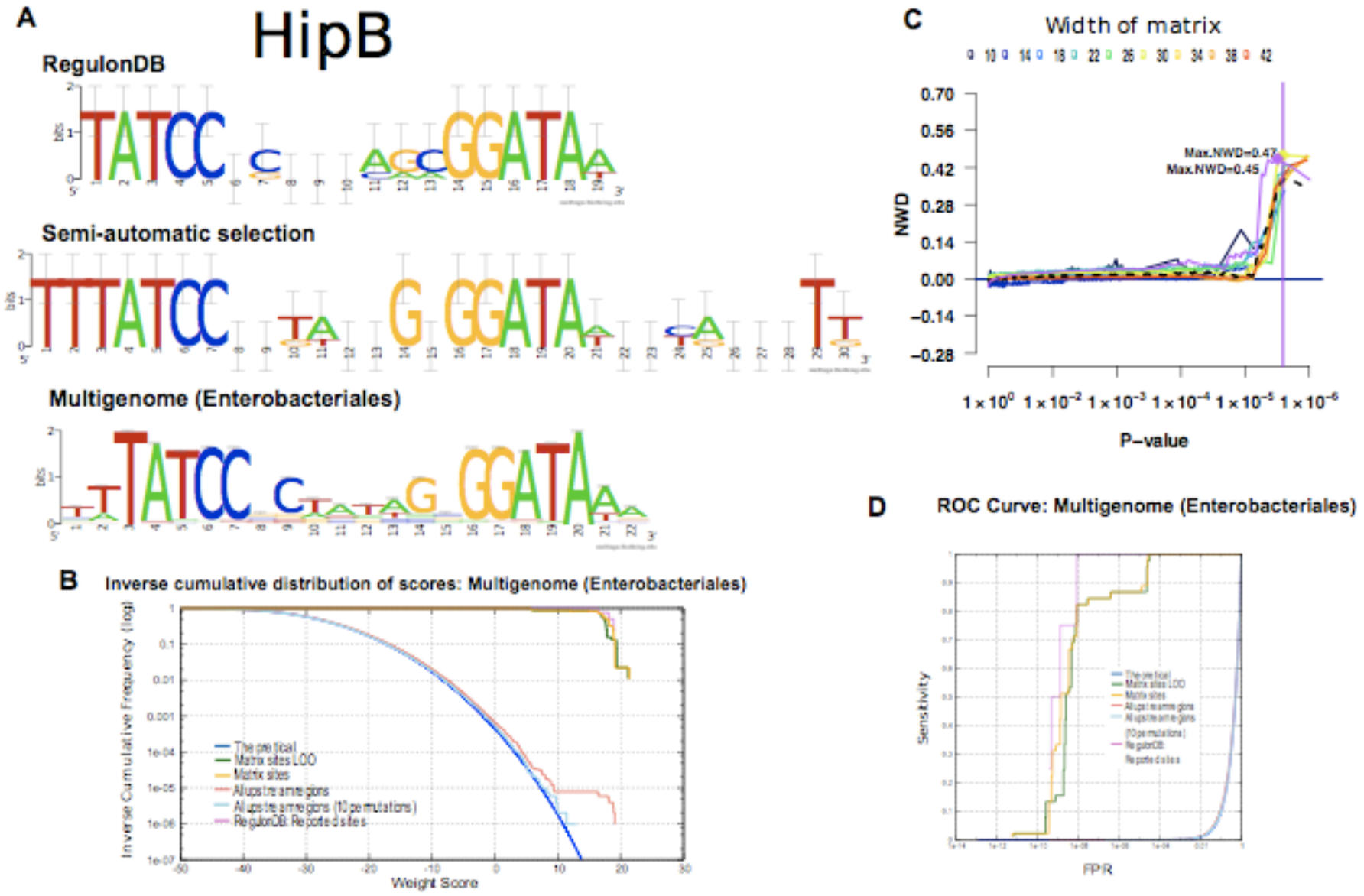






NanR matrix was built from 6 identical heptanucleotides.
The P-value distribution almost fits a binomial.
The LOO test is impossible (all sites are twins).

Using multigenome pattern discovery to enrich a matrix

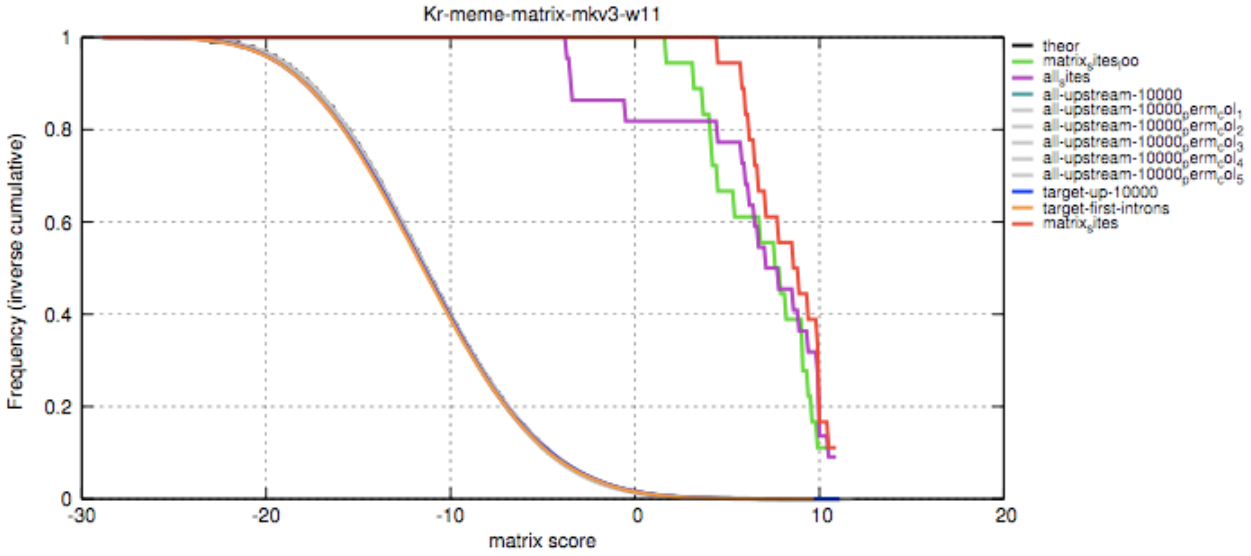


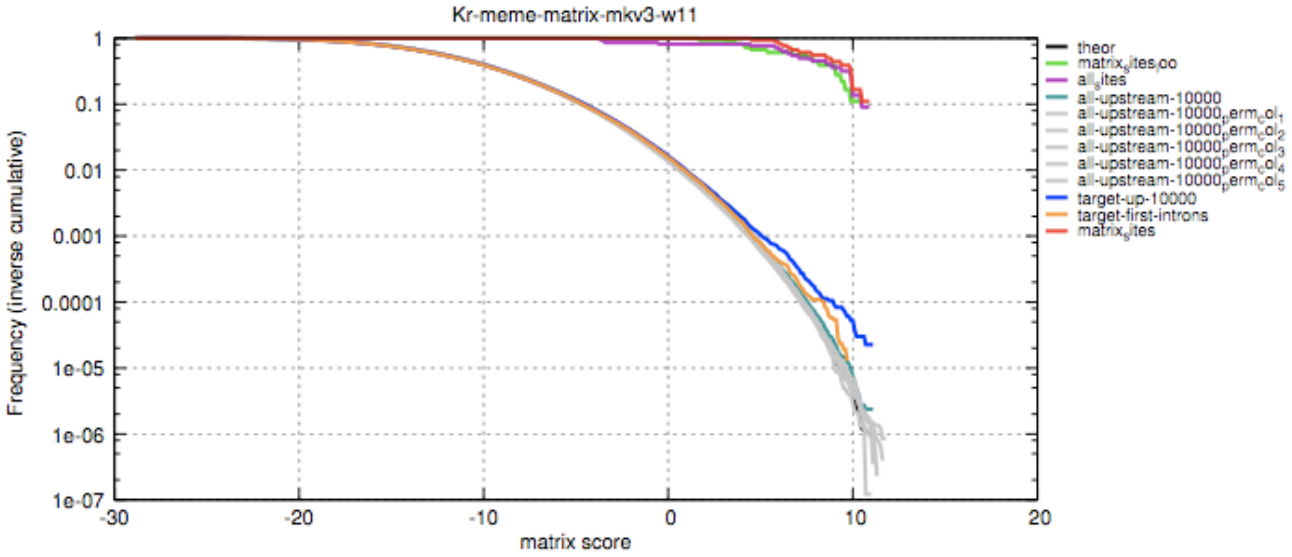
Selected study cases

Transcription factors from metazoans

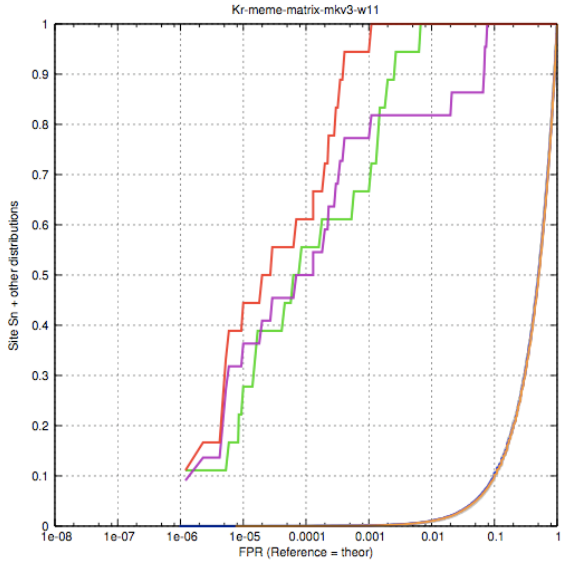
Jacques.van.Helden@ulb.ac.be
Université Libre de Bruxelles, Belgique
Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)
http://www.bigre.ulb.ac.be/

Matrix-quality for Drosophila Krüppel (Kr)

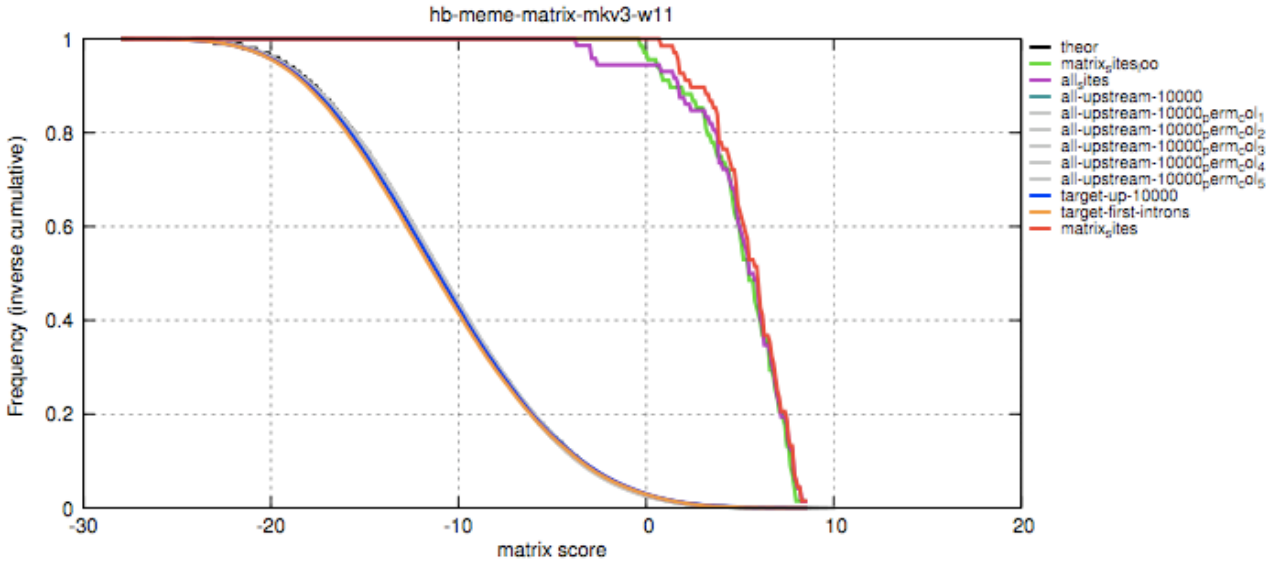


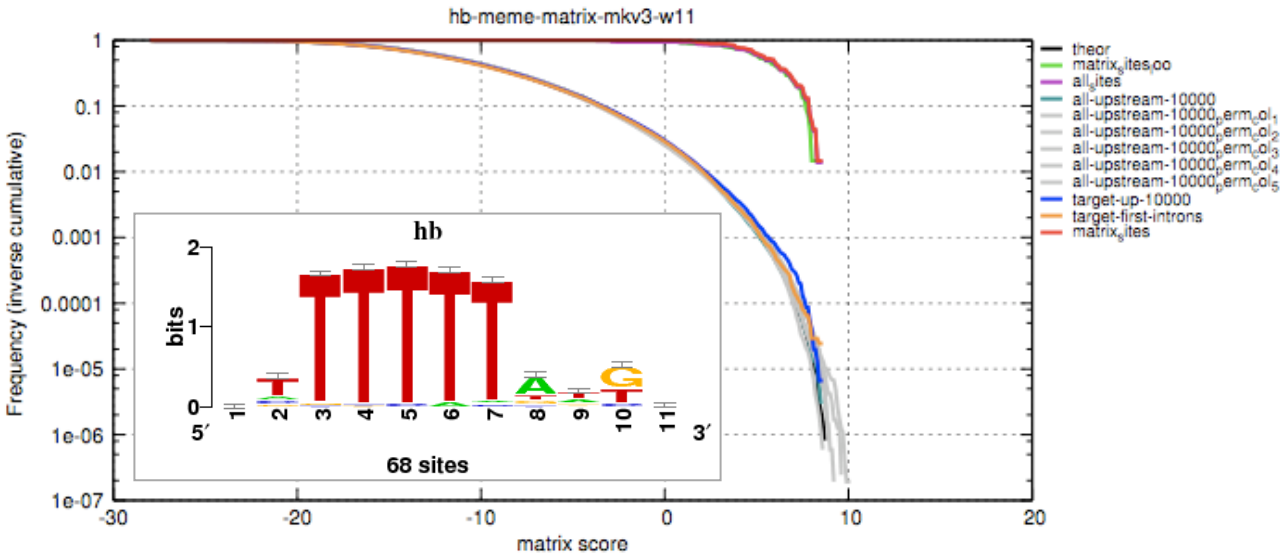


- Score distributions
  - Permutation tests fit theoretical distribution.
  - The upstream regions of Kr target genes show an enrichment in predicted sites (blue curve)
- ROC curve (green)
  - The $FPR_{0.50}$ is ~10^{-4}, indicating achieving a Sn of 50% costs 1FP every 10kb.
  - To achieve a Sn of 70%, the cost raises to 1FP/kb.

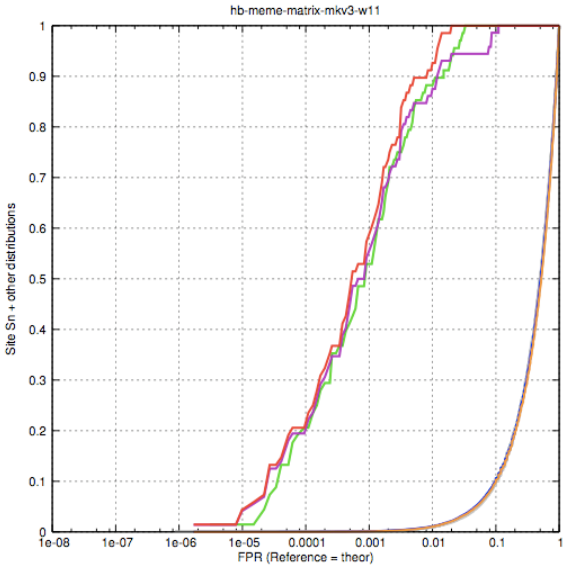


Matrix-quality for Drosophila Hunchback

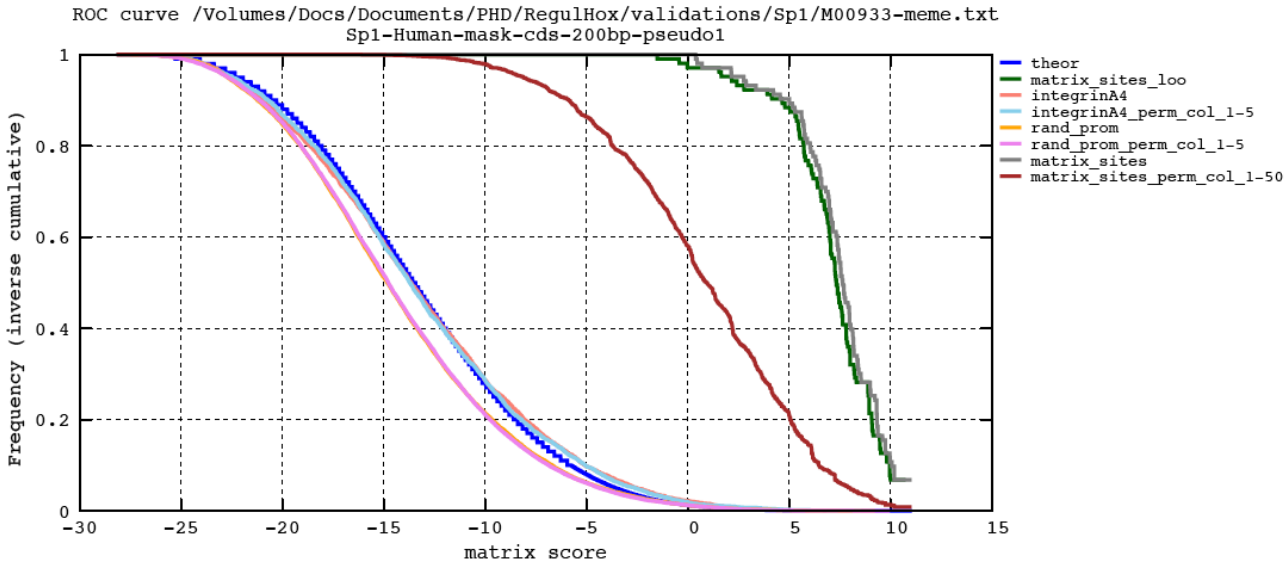




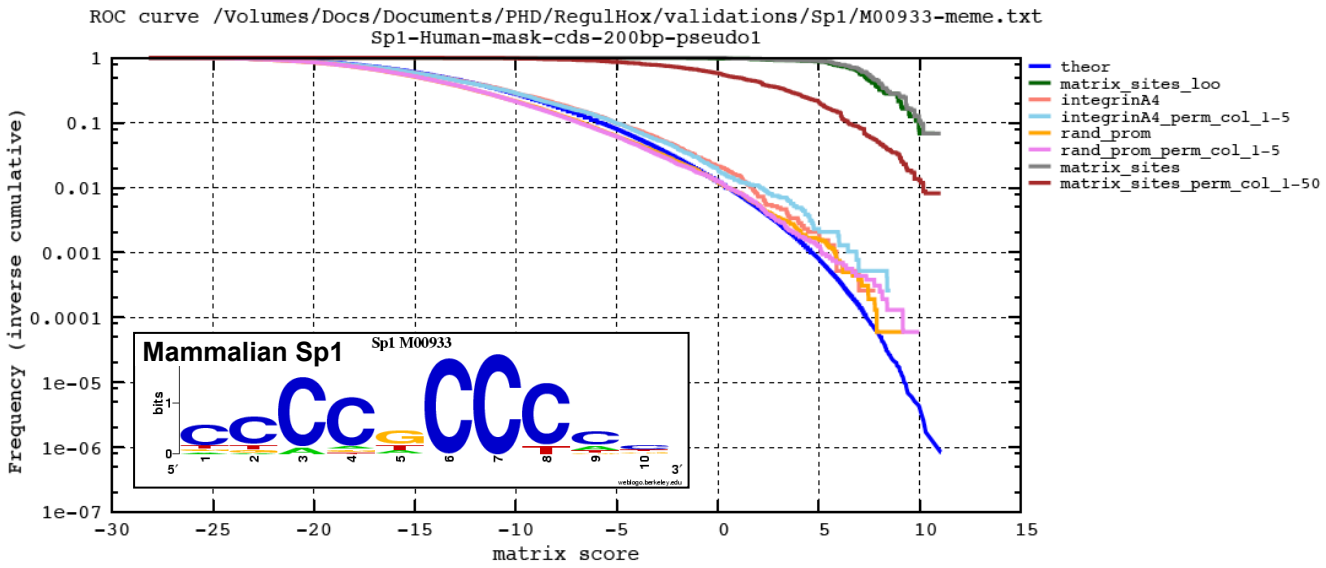
- The drosophila factor hunchback (hb) binds a T-repetitive motif.
- The motif is poorly informative, it is found ubiquitously in the genome.
- The $FPR_{0.50}$ is ~10^{-3}, indicating achieving a Sn of 50% costs 1FP every kb.
- The target genes (blue) show no particular enrichment relative to the theoretical distribution (black).

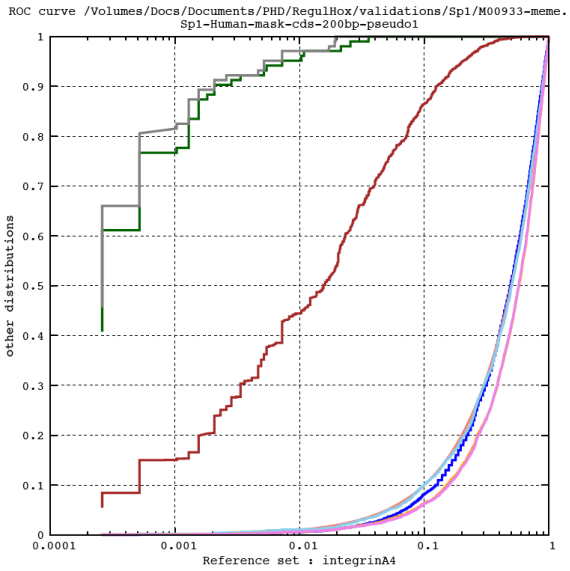


Matrix-quality result for Human Sp1



- The human transcription factor Sp1 binds a C-rich motif.
- The permuted motif is almost identical to the original motif.
- The matrix column-shuffling test is thus inoperant in this case.





Points for discussion

- Assessment based on binding sites
  - Is it possible to develop datasets with full characterization of binding elements?
  - Using artificial data (implanted motifs) ?
- Distribution-based assessment
  - One could consider to assess a motif on the basis of its whole distribution
  - Simple visual inspection already gives us informative clues
  - How to quantify this systematically ?
- Matrix parameters
  - Various parameters can be computed on a matrix
    - Information content
    - E-value of a matrix
    - MAP
  - Do these parameter give us reliable indication on the biological relevance of a motif ? On the specificity of the TF ?
  - Note: some of these parameters are used for optimizing matrices in pattern discovery (more details tomorrow)

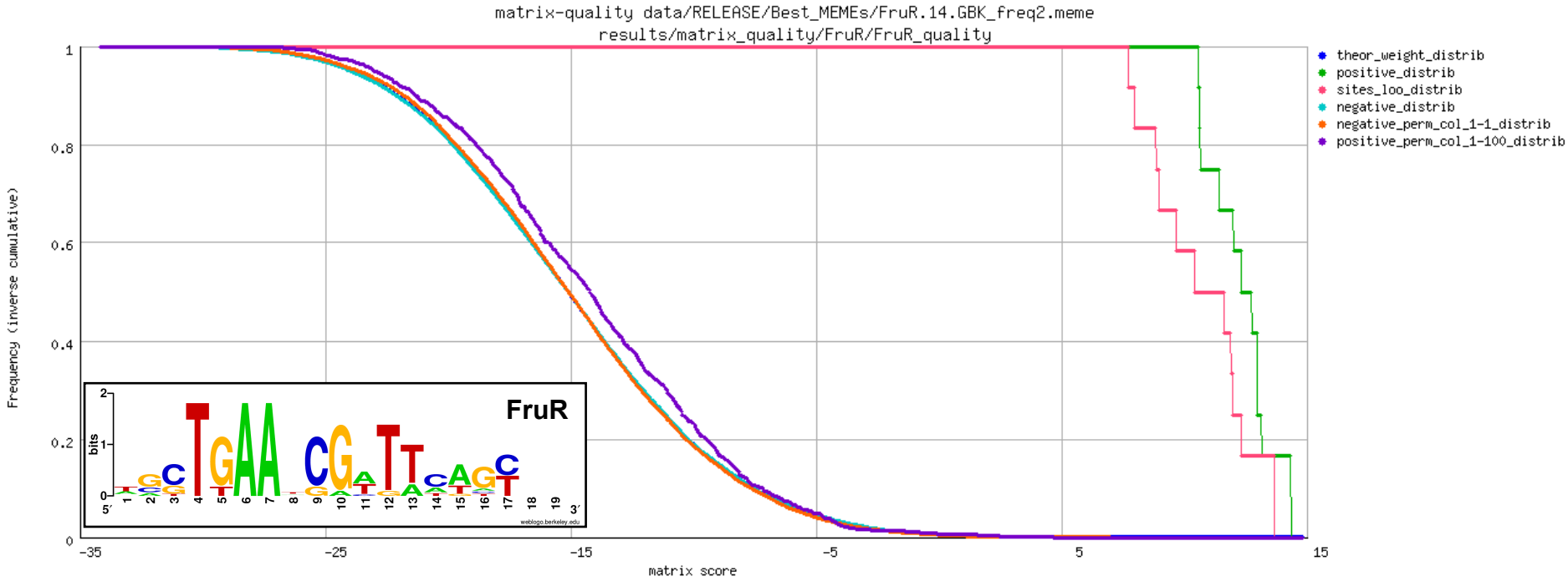
Regulatory Sequence ANalysis

Supplementary material

Jacques.van.Helden@ulb.ac.be
Université Libre de Bruxelles, Belgique
Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)
http://www.bigre.ulb.ac.be/

Estimating matrix quality
FruR from RegulonDB

- We compare the following inverse cumulative distributions of scores
  - Blue : theoretical distribution
  - Green: annotated binding sites that served to build the matrix (this is biased !)
  - Pink: leave-one-out (LOO) test with annotated binding sites
  - Violet: scores obtained with 100 permuted matrices in the annotated sites
  - Cyan: distribution in a random selection of promoters
  - Orange: scores obtained with 5 permuted matrices in a random selection of promoters

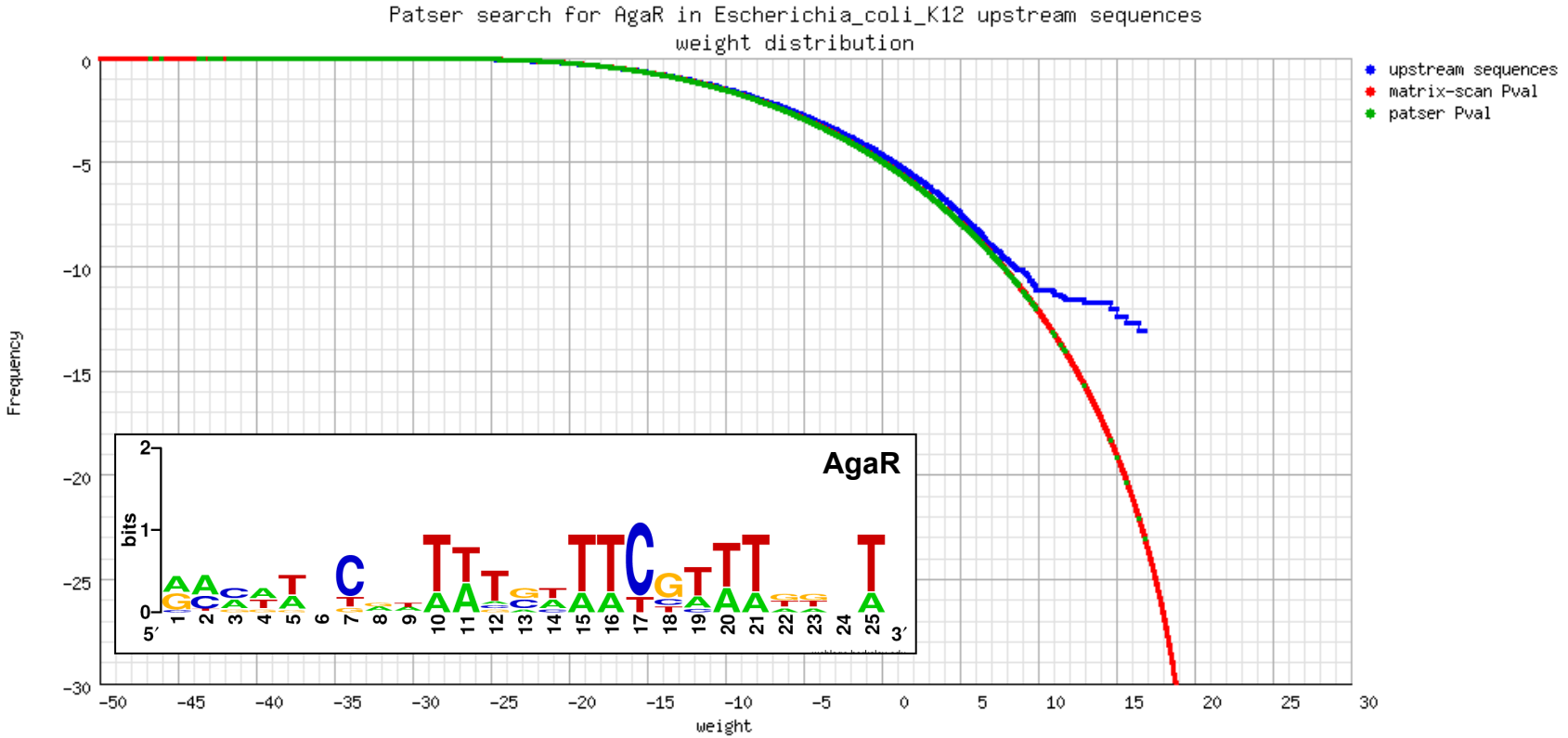


Escherichia coli PurR matrix

A	11	1	0	3	17	17	13	2	1	3	0	1	3	0	2	3	1
T	1	3	0	1	0	0	2	0	0	11	15	15	6	0	3	10	11
C	2	8	0	13	0	0	1	15	0	1	0	0	2	17	4	3	3
G	3	5	17	0	0	0	1	0	16	2	2	1	6	0	8	1	2
	G	C	G	C	A	A	G	C	G	T	T	T	T	C	C	A	G
	A	C	G	A	A	A	A	C	G	A	T	T	G	C	T	T	T
	A	C	G	C	A	A	C	C	G	T	T	T	T	C	C	T	T
	C	G	G	A	A	A	A	C	G	T	T	T	G	C	G	T	T
	A	G	G	A	A	A	A	C	G	T	T	T	C	C	G	C	T
	C	T	G	C	A	A	A	A	A	C	T	G	G	C	C	G	C
	A	G	G	C	A	A	A	C	G	T	T	T	A	C	C	T	T
	G	A	G	C	A	A	A	C	G	T	T	T	C	C	A	C	T
	A	C	G	C	A	A	A	C	G	T	T	T	T	C	T	T	T
	A	C	G	C	A	A	A	C	G	T	G	T	G	C	G	T	C
	T	C	G	C	A	A	A	C	G	T	T	T	G	C	T	T	T
	A	T	G	C	A	A	T	C	G	G	T	T	A	C	G	C	T
	A	G	G	T	A	A	T	C	G	T	T	T	G	C	G	T	A
	A	T	G	C	A	A	A	C	G	A	T	T	T	C	A	A	G
	A	C	G	C	A	A	A	C	G	G	T	T	T	C	G	T	C
	A	G	G	C	A	A	A	A	G	A	G	A	A	C	G	A	T
	G	C	G	C	A	A	A	C	G	T	T	T	T	C	G	T	T

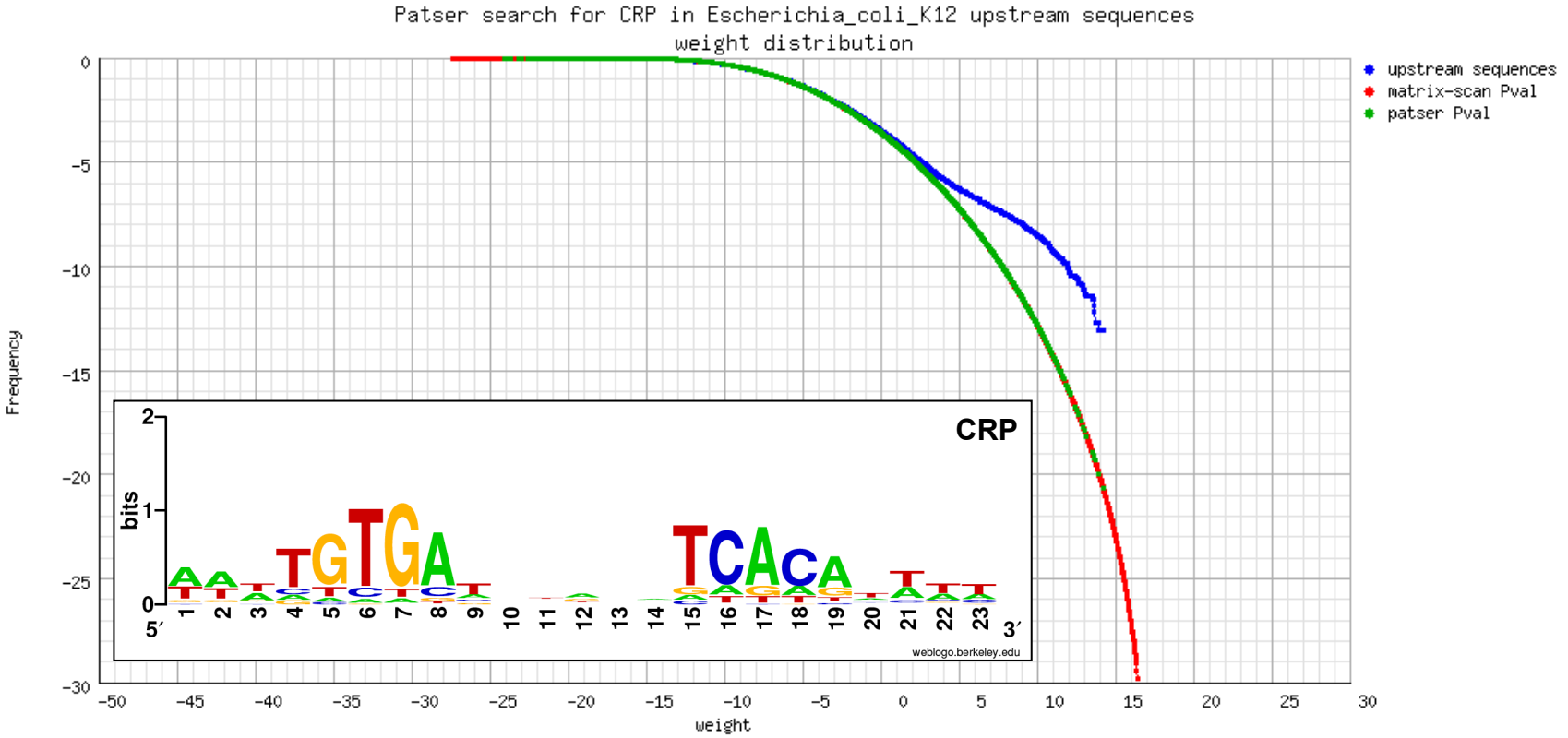
Distribution of matrix scores in E.coli promoters
AgaR matrix from RegulonDB

- We compare the distribution observed in all the E.coli promoters (blue) with the theoretical distribution, computed by patser (green) or matrix-scan (blue).
  - For low score values (W <= 5) observed distribution follows pretty well the theoretical one.
  - We also observe a few sites with a very high score(remind, the axis is logarithmic). These high scoring sites are more frequent than expected by chance.
  - These high-scoring sites hopefully correspond to the binding sites of ArgR, but we would of course like to check this.



Distribution of matrix scores in E.coli promoters
CRP matrix from RegulonDB

- CRP is a general transcription factor in E.coli
- RegulonDB contains annotations about ~130 target genes and ~200 binding sites
- Questions:
  - does the distribution reflect the genericity of the TF ?
  - Is this genericity related to
    - a weak specificity of the matrix ?
    - a large number of specific sites ?



Example of position-specific scoring matrix

- Binding sites and motif of the human cAMP-response element-binding protein (CREB)

Alignment of CREB binding sites (TRANSFAC annotations)

R01508	T	G	A	C	G	T	C	A	C	G
R15485	T	G	A	C	G	T	C	A	A	G
R04356	T	G	A	C	G	T	C	A	A	G
R16066	T	G	A	C	G	T	C	A	C	C
R00592	T	G	A	C	G	T	C	A	T	G
R14792	T	G	A	T	G	T	C	A	C	T
R12540	T	A	A	C	G	T	C	A	C	A
R04029	T	G	A	C	A	T	C	A	C	G
R16774	G	T	A	C	G	T	C	A	C	G
R02906	G	T	A	C	G	T	C	A	C	G
R09519	T	G	A	C	G	T	C	C	A	T
R14826	G	C	A	C	G	T	C	A	A	G
R16161	T	G	A	C	G	A	C	A	A	C
R14781	T	C	A	C	G	T	A	A	C	T
R17209	T	G	A	C	G	C	T	A	C	G
R08102	T	G	A	G	C	T	C	A	C	T
R04963	T	G	A	C	G	T	C	T	G	A
R12331	A	A	A	C	G	T	C	A	T	C
R14791	A	A	A	T	G	T	C	A	C	A

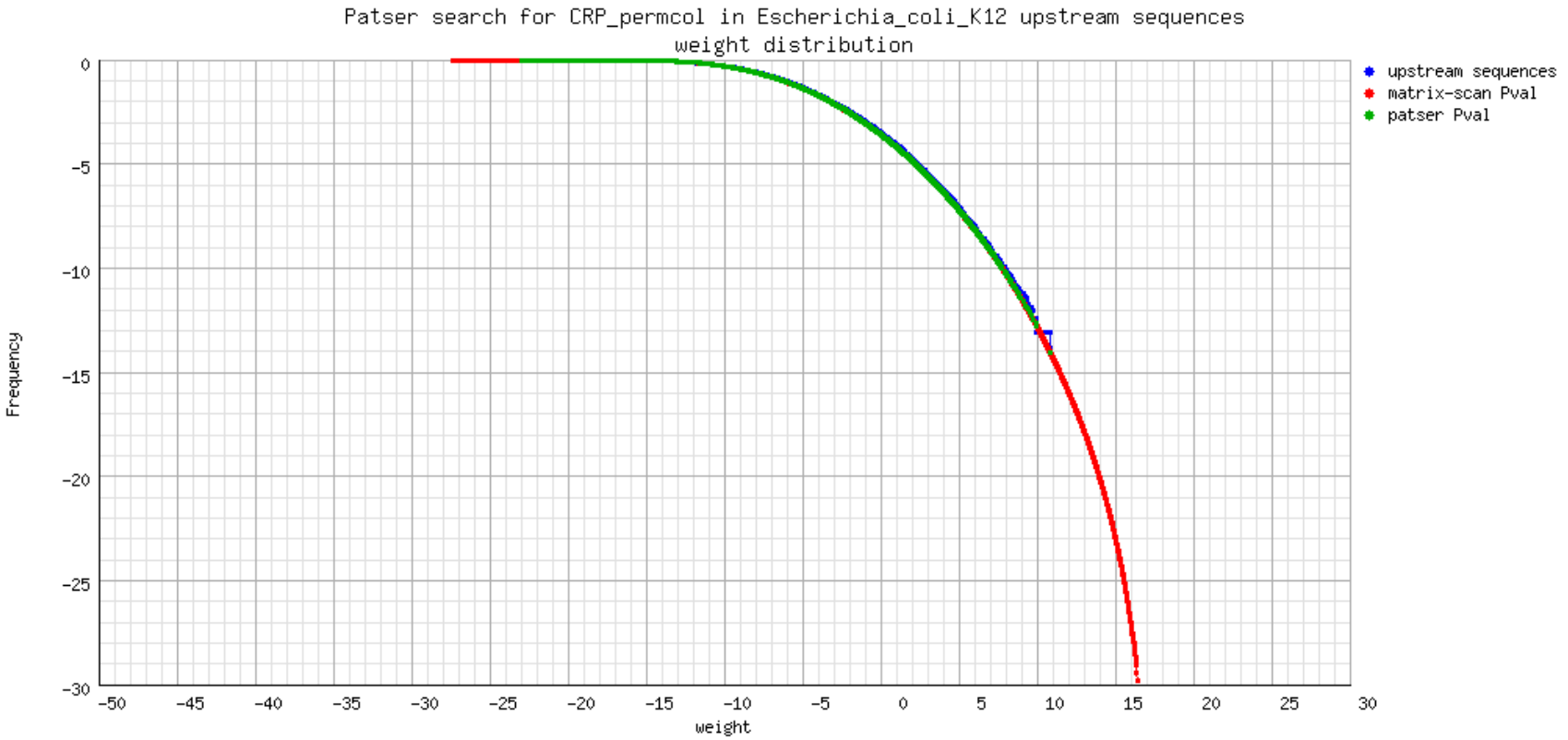
Count matrix (TRANSFAC matrix CREB)

residue\position	1	2	3	4	5	6	7	8	9	10
A	2	3	**19**	0	1	1	1	**17**	5	3
C	0	2	**0**	**16**	1	1	**17**	1	**11**	3
G	3	**12**	0	**1**	**17**	0	0	0	1	**9**
T	**14**	2	0	2	0	**17**	1	1	2	4
Sum	19	19	19	19	19	19	19	19	19	19

Figure from
Jean Valéry Turatsinze

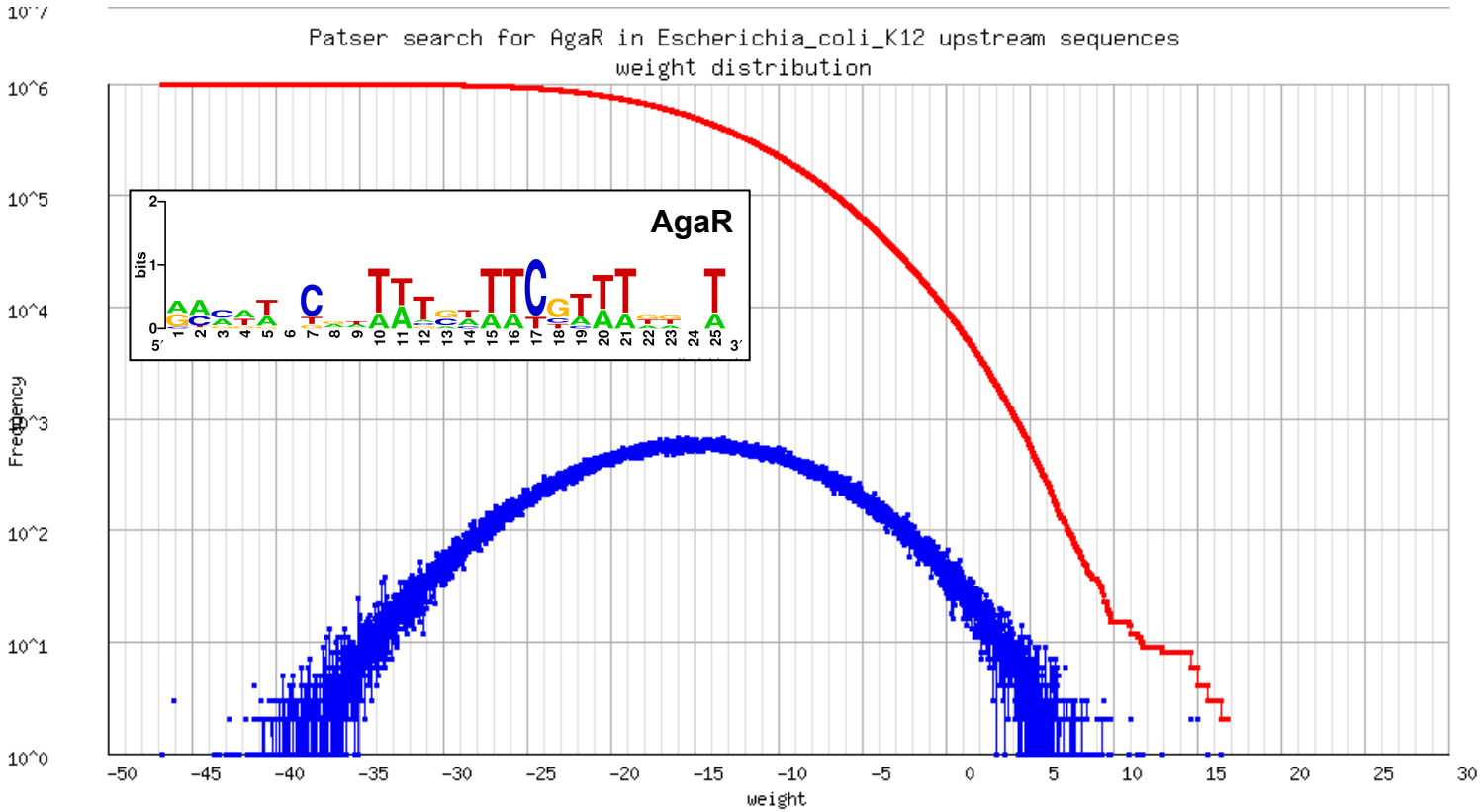
Distribution of matrix scores in E.coli promoters
CRP matrix from RegulonDB - permuted

- CRP is a general transcription factor in E.coli
- RegulonDB contains annotations about ~130 target genes and ~200 binding sites
- Questions:
  - does the distribution reflect the genericity of the TF ?
  - Is this genericity related to
    - a weak specificity of the matrix ?
    - a large number of specific sites ?

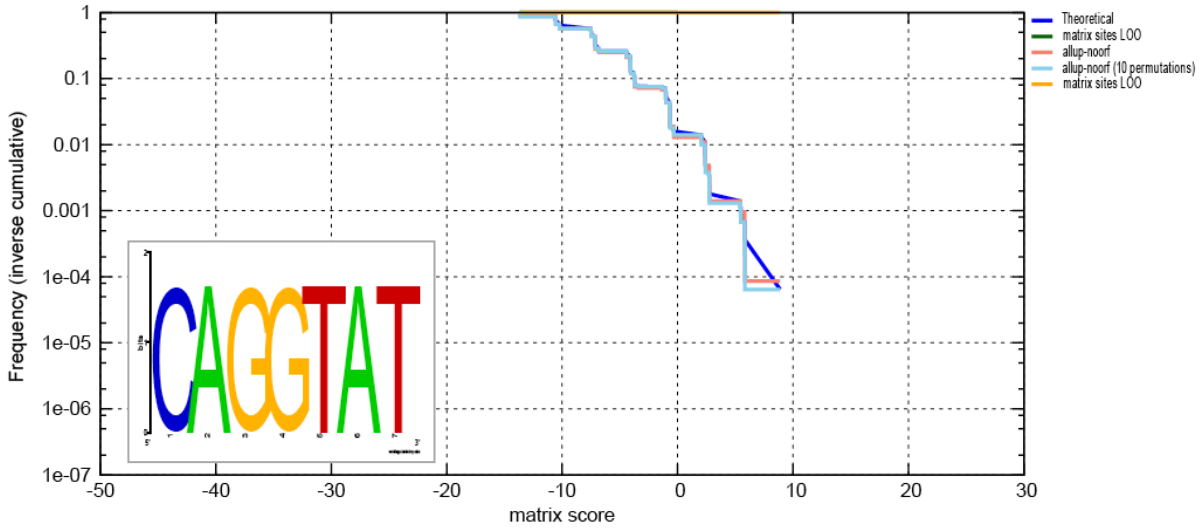


Distribution of matrix scores in E.coli promoters

- We scanned upstream sequences for all the genes of *Escherichia coli K12*, with each matrix of RegulonDB.
- This is illustrated below with the distribution of scores assigned with the matrix AgaR for all the positions in the complete collection of *E.coli* promoters.
  - Blue: distribution of scores in all promoters (beware: Y axis is logarithmic)
  - Red: inverse cumulative distribution.

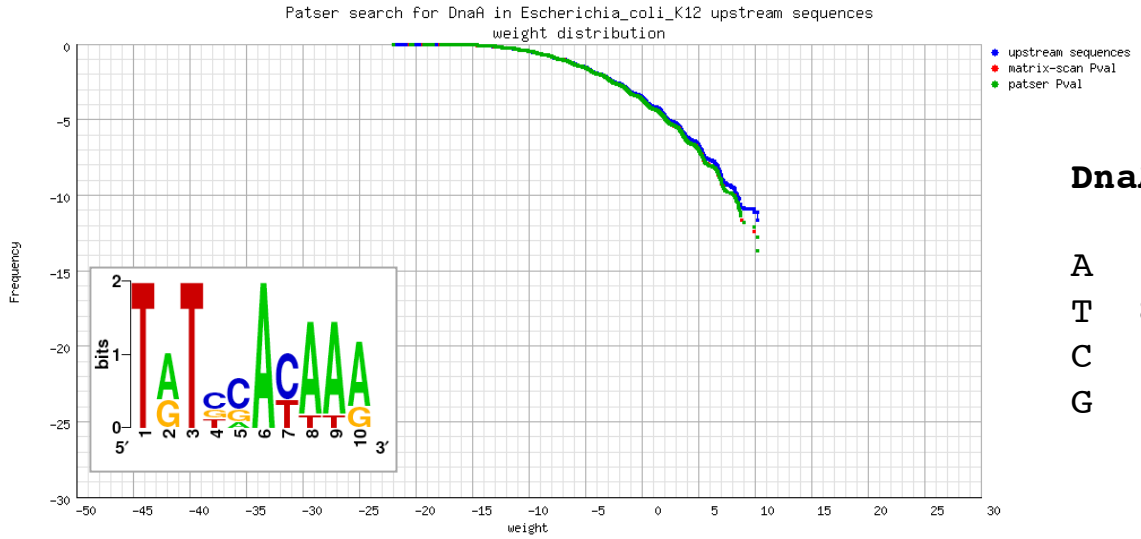


Some tricky cases



NanR matrix

A	0	**6**	0	0	0	**6**	0
T	0	0	0	0	**6**	0	**6**
C	**6**	0	0	0	0	0	0
G	0	0	**6**	**6**	0	0	0



DnaA matrix

A	0	5	0	0	1	**8**	0	**7**	**7**	**6**
T	**8**	0	**8**	2	0	0	3	1	1	0
C	0	0	0	4	5	0	5	0	0	0
G	0	3	0	2	2	0	0	0	0	2

Theory

Otto G. Berg† and Peter H. von Hippel

Institute of Molecular Biology
University of Oregon
Eugene, Oregon 97403, U.S.A.

(Received 7 May 1986, and in revised form 29 September 1986)

We present a statistical-mechanical selection theory for the sequence analysis of a set of specific DNA regulatory sites that makes it possible to predict the relationship between individual base-pair choices in the site and specific activity (affinity). The theory is based on the assumption that specific DNA sequences have been selected to conform to some

- In 1987, Otto G. Berg and Peter H. von Hippel propose a theoretical model to predict binding affinity between regulatory proteins and their DNA binding sites.
- They compare binding affinity and score

Quality of the CRP matrix