



## TRANSCRIPTION FACTOR WEIGHT MATRIX EVALUATION

Alejandra Eugenia Medina Rivera

RegulonDB database



Copies and Copyright-Notice

User is not entitled to copy or redistribute parts of the RegulonDB database or to eliminate, copy or modify it.  
User is not entitled to copy or redistribute parts of the RegulonDB database or as a whole into other databases without prior written permission from CCG-UNAM.

Citation

User is committed to cite properly the work of the RegulonDB team and update publication concerning RegulonDB:  
Socorro Gama-Castro, Verónica Jiménez-Jacinto, Pedro A. Salgado, Alberto Santos-Zavaleta, Mónica Martínez-Flores, Heladio Moreira, Juan Segura-Salazar, Luis Muñiz-Rascado, Irma Martínez-Flores, Heladio Salgado, Carlos Bonavides-Martínez, Cei Almouzni, José Carlos Rodríguez-Pousa, Juan Miranda-Rios, Enrique Morett, Enrique Merino, Alfonso Valencia and Julio Collado-Vides.  
"RegulonDB (Version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and transcription navigation"  
Nucleic Acids Research, 2008; Vol. 36, 2320-2324.  
Release: 6.4 Date: 10-AUG-09

This presentation aims to show the user the methodology followed for the evaluation of RegulonDB matrices. It presents the statistic and concepts involved.

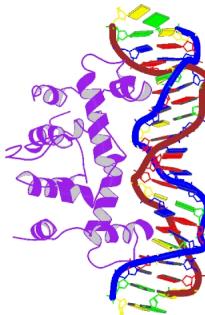
## Summary:

- Building model to predict TFBSS
- Matrix-quality program.
- Matrix Score distribution
- Evaluation criteria

Copies and Copyright-Notice

User is not entitled to change or erase data of the databases or to eliminate copyright notices from RegulonDB. User is not entitled to expand RegulonDB or to integrate or as a whole into other databank systems, without prior permission from CCG-UNAM.

Citation



## A Annotated TrpR binding sites

### Site ID

Site ID	G	T	A	C	T	A	G	T	T	T	G	A	T	G	G	T	A	T	G
ECK120012644	G	T	A	C	T	A	G	T	T	T	G	A	T	G	G	T	A	T	G
ECK120012187	G	T	A	C	T	A	G	T	T	T	G	A	T	G	G	T	A	T	G
ECK120012179	G	A	A	C	T	A	G	T	T	A	A	C	T	A	G	T	A	C	G
ECK120012892	G	A	A	C	T	A	G	T	T	A	A	C	T	A	G	T	A	C	G
ECK120012181	G	A	A	C	T	A	G	T	T	A	A	C	T	A	G	T	A	C	G
ECK120012636	G	T	A	C	T	A	G	A	G	A	A	C	T	A	G	T	G	C	A
ECK120012183	G	T	A	C	T	A	G	A	G	A	A	C	T	A	G	T	G	C	A
ECK120012185	G	T	A	C	T	C	G	T	G	T	A	C	T	G	G	T	A	C	A
ECK120012979	G	T	A	C	T	C	G	T	G	T	A	C	T	G	G	T	A	C	A
ECK120012894	G	T	A	C	T	C	T	T	T	A	G	C	G	A	G	T	A	C	A

### Target Operon

aroL-yaiA-aroM  
aroL-yaiA-aroM  
trpLEDCBA  
trpLEDCBA  
trpLEDCBA  
trpLEDCBA  
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	8	0	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



# Pattern-matching: scanning sequences for putative TFBSS

Copies and Copyright

User is not entitled to  
distribute or to eliminate  
User is not entitled to  
or as a whole into other  
from CCG-UNAM.

Citation

User is committed to  
update publication of  
Socorro Gama-Castañ  
Alberto Santos-Zava  
Moreira,  
Juan Segura-Salazar  
Salgado  
César Bonavides-Ma  
Juan Miranda-Rosa  
Enrique Morett, Enri  
and Julio Collado-Vi  
"RegulonDB (Version  
beyond transcription  
textpresso navigation  
Nucleic Acids Resea

Release: 6.4 Date: 1



Transcription Factor

The consensus and  
Transcription Factor

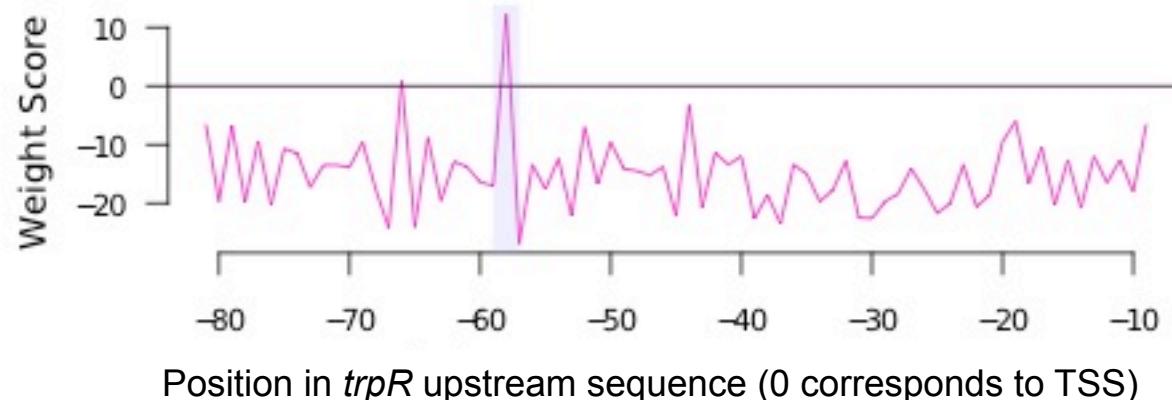
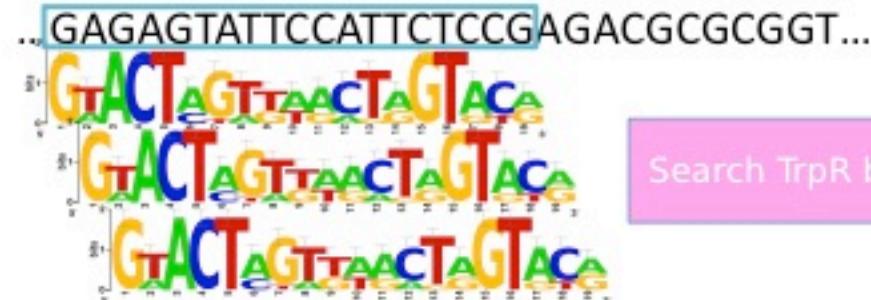
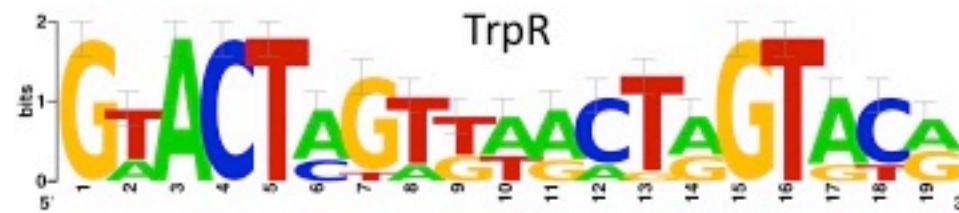
Total of uniq bindin

Matrix

A	2	3	1	1	3
C	1	0	0	1	0
G	0	0	1	0	0
T	0	0	1	1	0

AlignmentScore

```
AAGCAAAGGCCA
AAAAAAATTAAAG
CATTACATTGGCTG
```



# Matrix quality

## Copies and Copyright-Notice

User is not entitled to change or erase data sets of the RegulonDB databases or to eliminate copyright notices from RegulonDB. Furthermore, User is not entitled to expand RegulonDB or to integrate RegulonDB partly or as a whole into other databank systems, without prior written consent from CCG-UNAM.

## Citation

User is committed to cite properly the use of RegulonDB. The current most update publication concerning RegulonDB is:  
 Socorro Gama-Castro, Verónica Jiménez-Jacinto, Martín Peralta-Gámez, Alberto Santos-Zavaleta, Montserrat Salgado, Ana M. García-Salguero, Juan Segura-Salazar, Luis Muñoz-Alonso, Carlos Bonavides-Martínez, Cecilia Gómez, Juan Miranda-Rios, Enrique Morett, Enrique Merino, and Julio Collado-Vides.  
 "RegulonDB (Version 6.0): genome-wide transcriptional regulation beyond transcription, active (experimental) promoters and expression textpresso navigation." Nucleic Acids Research, 2008; vol. 36, D42-D44.  
 Release: 6.4 Date: 10-AUG-09

## Transcription Factor Matrix and Alignments

The consensus and patser programs were used to create the matrix and alignment  
 Transcription Factor Name: Ada  
 Total of uniq binding sites: 3

Matrix

A	2	3	1	1	3	2	3	0	1	1	1	1	2	0	0	2	0	3	2	0	1	1	0	0	1	0	1	0
C	1	0	0	1	0	1	0	0	1	0	2	0	2	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0
G	0	0	1	0	0	0	1	0	2	0	3	1	2	0	0	0	3	0	0	2	0	1	0	1	0	0	0	0
T	0	0	1	1	0	0	2	2	0	1	0	0	2	0	1	0	0	1	3	0	2	2	2	2	2	2	2	2

AlignmentScore  
 AAGCAAAGGCCAGCGTCTGAATAACGTTT20.66  
 AAAAAAATTAAGGGCAAGATTGTTGGTT21.42  
 CATTACATTGCTGGATAAGAATGTTCAGTAG19.78

## Evaluation

- Is the matrix **good to predict new putative binding sites?**
- Which is the **sensitivity** to recover **true binding sites**?
- Which is the **false positive rate** for a given **sensitivity**?

<sup>1</sup>Bailey and Elkan. Systems for Molecular Biology (1994)

<sup>2</sup>Schneider et al. J Mol Biol (1986)

#### Copies and Copyright-Notice

User is not entitled to change or erase data sets of the RegulonDB databases or to eliminate copyright notices from RegulonDB. Furthermore, User is not entitled to expand RegulonDB or to integrate RegulonDB partly or as a whole into other databank systems, without prior written consent from CCG-UNAM.

#### Citation

User is committed to cite properly the use of RegulonDB. The current most update publication concerning RegulonDB:  
 Socorro Gama-Castro, Verónica Jiménez-Jacinto, M. González-Gálvez, Alberto Santos-Zavaleta, Monika Collado-Vides, Heladia Salgado, Juan Segura-Salazar, Luis Muñiz-Ramírez, Irma Martínez-Pérez, Heladia Salgado, César Bonavides-Martínez, Cei Abreu-Goodger, Carlos Rodríguez-Penagos, Juan Miranda-Rios, Enrique Morett, Enric Merino, Áracci M. Huerta, Luis Treviño-Quintanilla and Julio Collado-Vides.  
 "RegulonDB (Version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription: active (experimental) annotated promoters and transcription navigation"  
*Nucleic Acids Research*, 2008; vol. 36. D740-D744.

Release: 6.4 Date: 10-AUG-09

#### Transcription Factor Matrix and Alignments

The consensus and patser programs were used to create Transcription Factor Name/DA Total of uniq binding sites3

Matrix

A	2	3	1	1	3	2	3	0	0	1	1	2	0	0	1	0
C	1	0	0	1	0	1	0	0	1	0	2	0	0	2	0	
G	0	0	1	0	0	0	0	1	0	2	0	0	3	1	2	
T	0	0	1	1	0	0	2	2	0	0	1	0	0	0	2	

AlignmentScore

```
AAGCAAAGCCACCGCTCTGAATAACGTTT20.66
AAAAAAATTAAAGGGCAAGATGTGGTT21.42
CATTACATTGCTGGATAAGAATGTTTAG19.78
```

**Nucleic Acids Research Advance Access published October 4, 2010**

*Nucleic Acids Research*, 2010, 1–17  
 doi:10.1093/nar/gkq710

## Theoretical and empirical quality assessment of transcription factor-binding motifs

**Alejandra Medina-Rivera<sup>1,2,\*</sup>, Cei Abreu-Goodger<sup>3</sup>, Morgane Thomas-Chollier<sup>4</sup>, Heladia Salgado<sup>1</sup>, Julio Collado-Vides<sup>1</sup> and Jacques van Helden<sup>1,2</sup>**

<sup>1</sup>Centro de Ciencias Genomicas, Universidad Nacional Autónoma de México. Av. Universidad s/n. Cuernavaca, Col. Chamilpa, Morelos 62210; Mexico, <sup>2</sup>Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe). Université Libre de Bruxelles, Campus Plaine, CP 263. Bld du Triomphe. B-1050 Bruxelles, Belgium,

<sup>3</sup>EMBL—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK and <sup>4</sup>Department of Computational Molecular Biology. Max Planck Institute for Molecular Genetics.

Ihnestrasse 73. 14195 Berlin, Germany

Received February 11, 2010; Revised July 2, 2010; Accepted July 27, 2010

Medina-Rivera *et al.* Nucleic Acids Research (2010) pp.

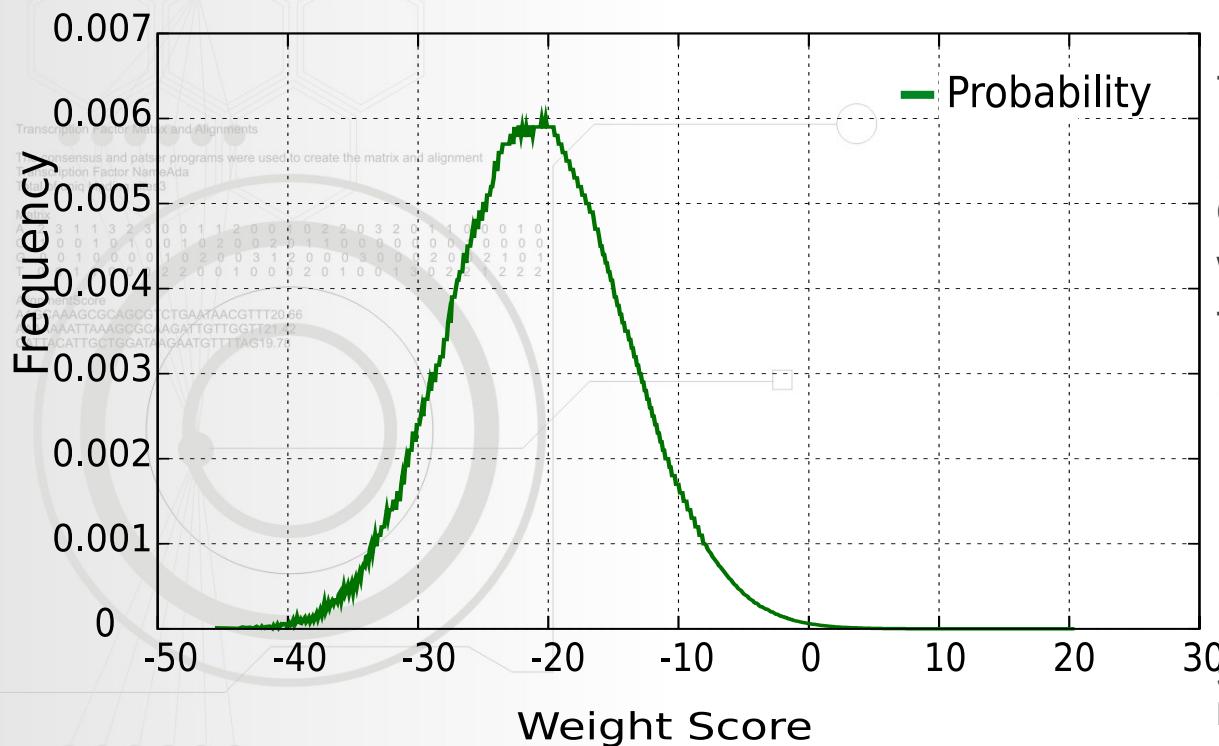
# Which are all the possible *scores* that could be generated by a PSSM?

## Score distribution: Theoretical distribution

A	-0.06	0	1.08	-0.06	0.06	0.54	-0.06	-0.07	-0.06	0.38	0.54	-0.07	0.06	0.38	-0.06	-0.06	0.71	-0.06	0.24
T	-0.07	0.46	-0.07	-0.07	0.97	-0.07	-0.12	0.62	0.31	-0.06	-0.07	0.79	-0.07	-0.07	-0.07	0.97	0.07	-0.1	-0.07
C	-0.04	-0.04	-0.04	1.56	-0.04	0.15	-0.04	-0.04	-0.04	-0.04	-0.04	1.09	-0.04	-0.04	-0.04	-0.04	1.09	-0.04	0.04
G	1.42	-0.04	-0.04	-0.04	-0.04	-0.04	1.2	-0.04	0.25	-0.04	0.11	-0.04	-0.04	0.25	1.42	-0.04	0	-0.04	0.41

ATATACGTATCTACTACTTG =  
3.25

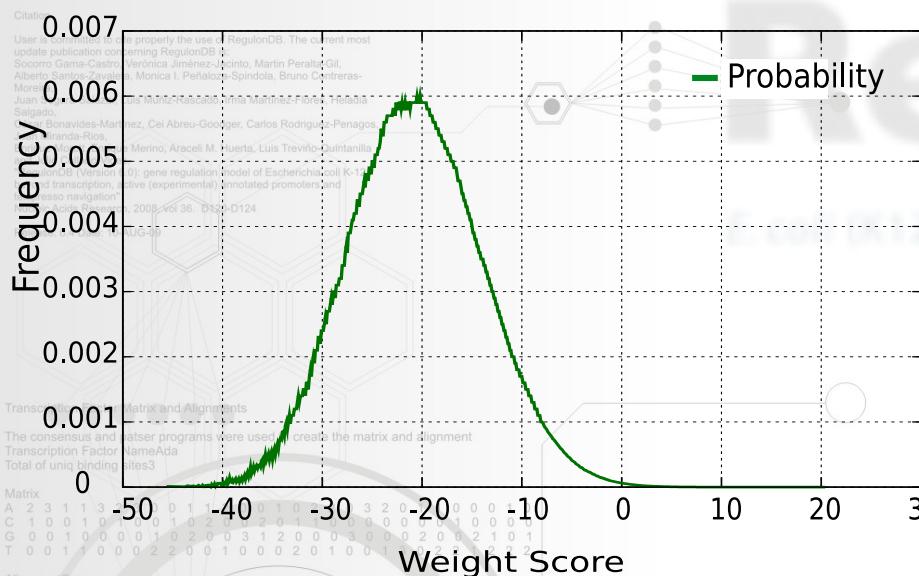
Theoretical probability of scores



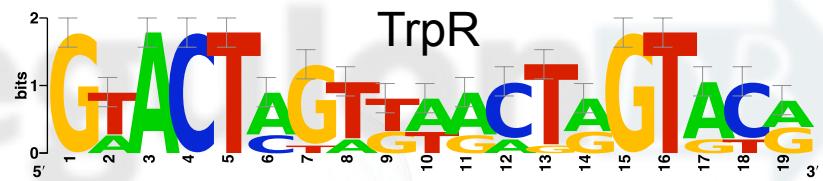
The ‘theoretical distribution’ provides an estimate of the expected FPR at each possible weight score (WS), based on the prior choice of a relevant background model

Staden. Comput Appl Biosci (1989)  
Extended to higher markov models : *matrix-distrib* (RSAT)

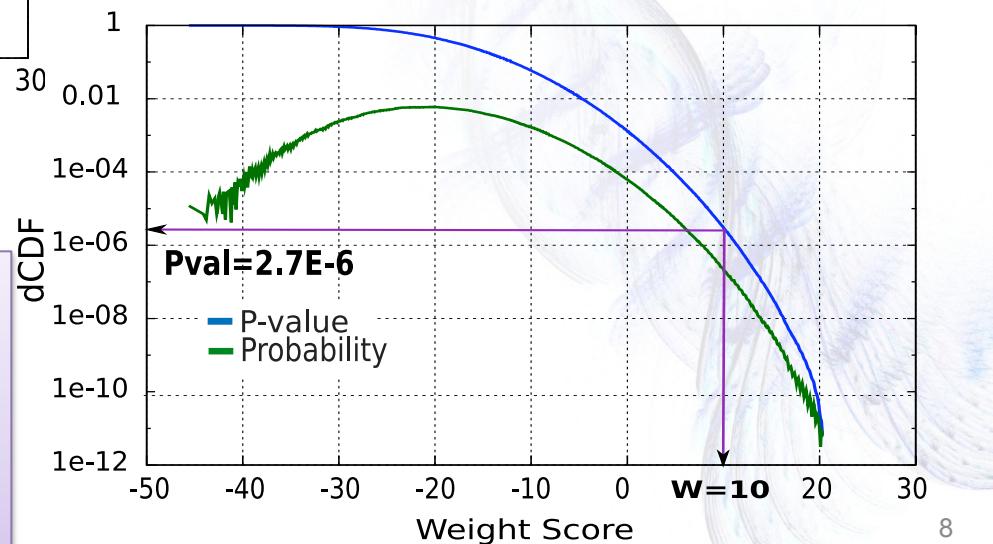
## Theoretical probability of scores



- E-value = P-value \* number of tested positions
  - E.G. When scanning all *E. coli* K12 upstream regions (L=603945 bps)
  - Pval =  $P(\text{ weight} \geq 10) = 2.7\text{e-}6$  (purple line)
  - Eval = Pval \* L \* 2  
 $= 2.7\text{e-}6 \text{ FP/bps} * 603945 \text{ bps} * 2 \sim 3$   
expected False Positives

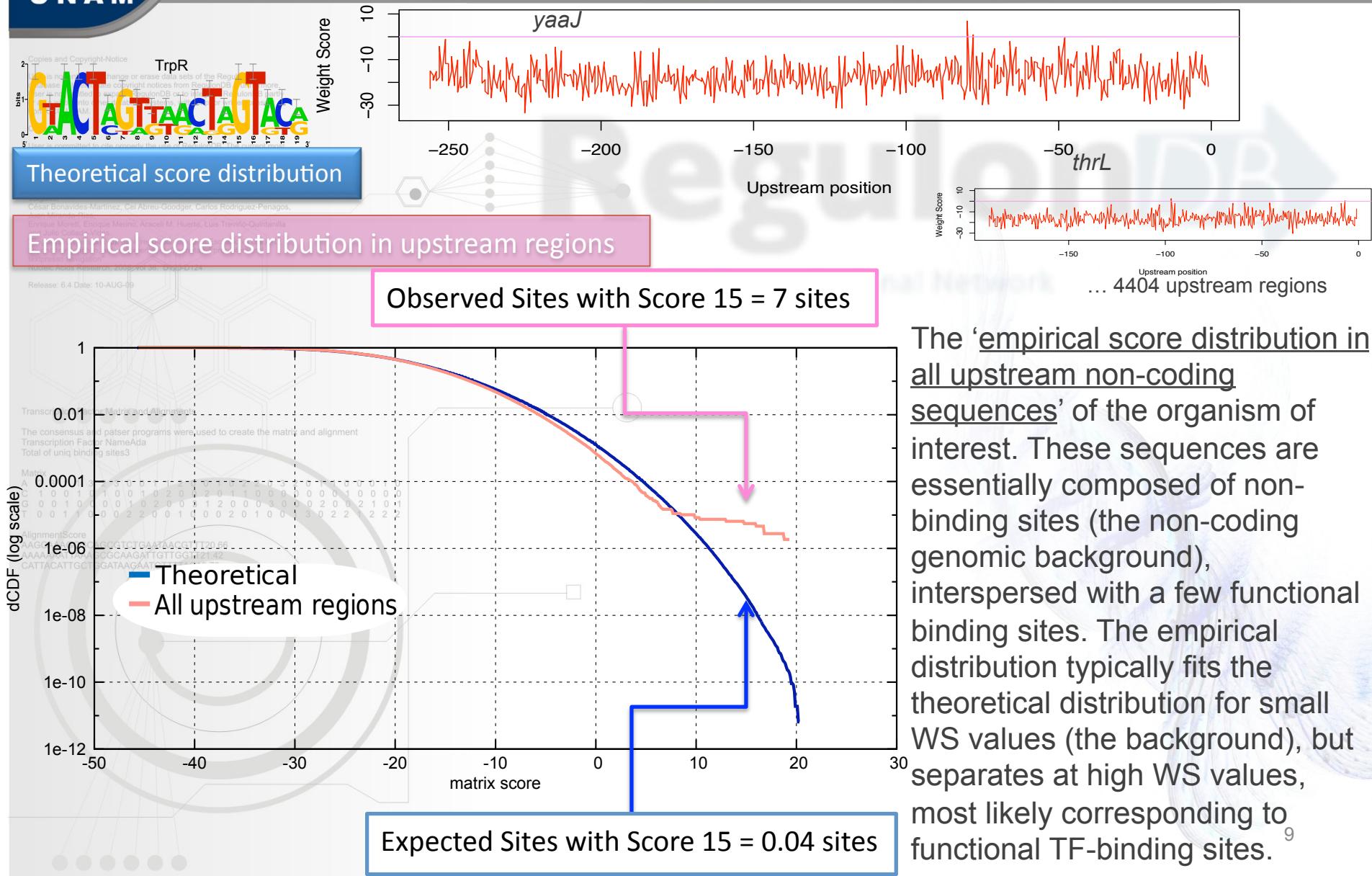


## Theoretical score distribution

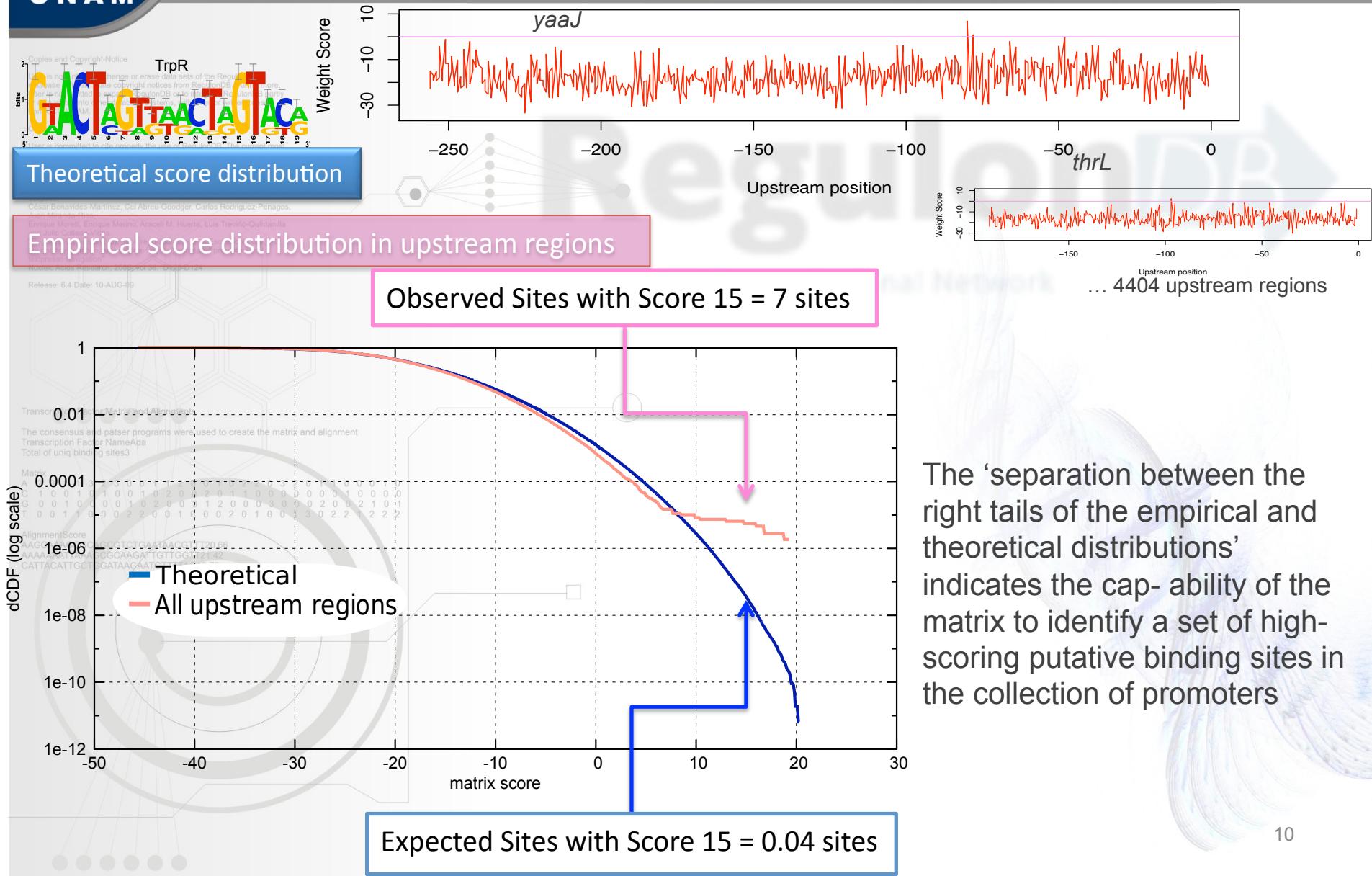


Staden. Comput Appl Biosci (1989) vol. 5 (4) pp. 293-8

# Empirical score distribution in all upstream non-coding sequences



# Empirical score distribution in all upstream non-coding sequences

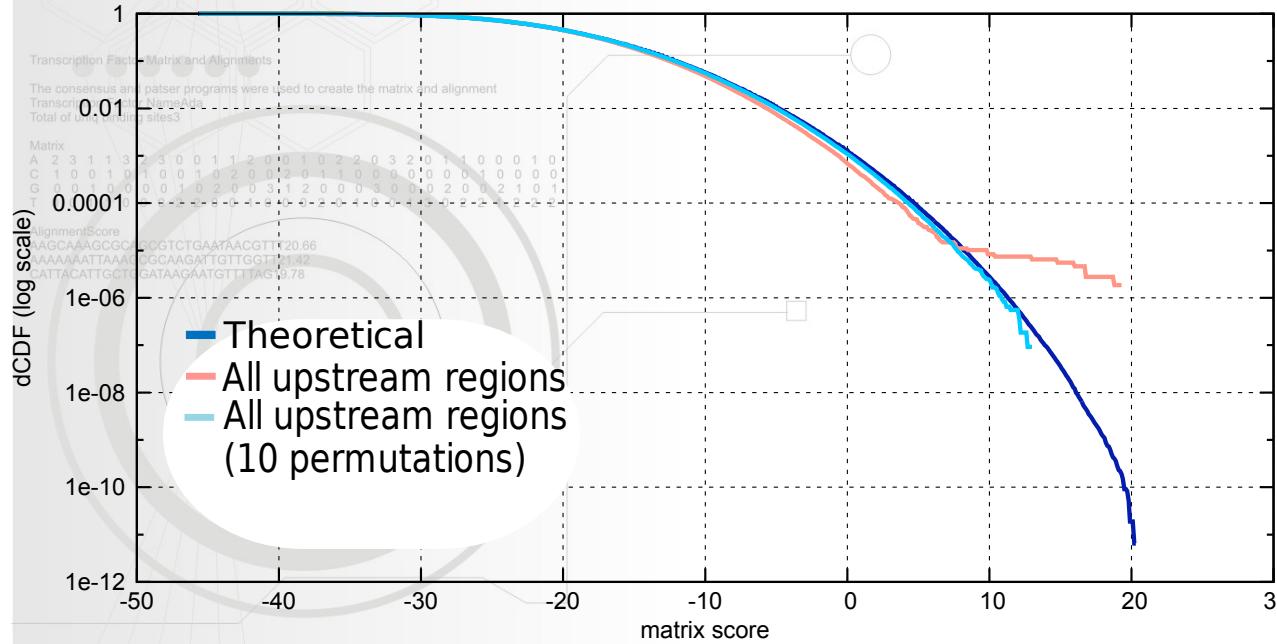


# Empirical score distribution in all upstream non-coding sequences

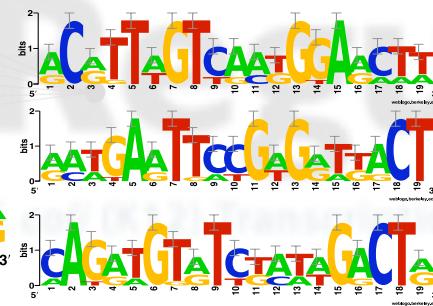
## Negative control: Permuted matrices

### Theoretical score distribution

Negative control: Permuted matrix.



### Empirical score distribution in upstream regions



An empirical estimate of the FPR is obtained by scanning all upstream non-coding sequences with column-permuted matrices, which supposedly do not correspond to any TF in the organism under consideration. If the background model has been chosen correctly, the ‘empirical distribution of the permuted matrices’ should fit the theoretical distribution

# Empirical score distribution in the annotated binding sites

Copies and Copyright-Notice

## Theoretical score distribution

Negative control: Permuted matrix.

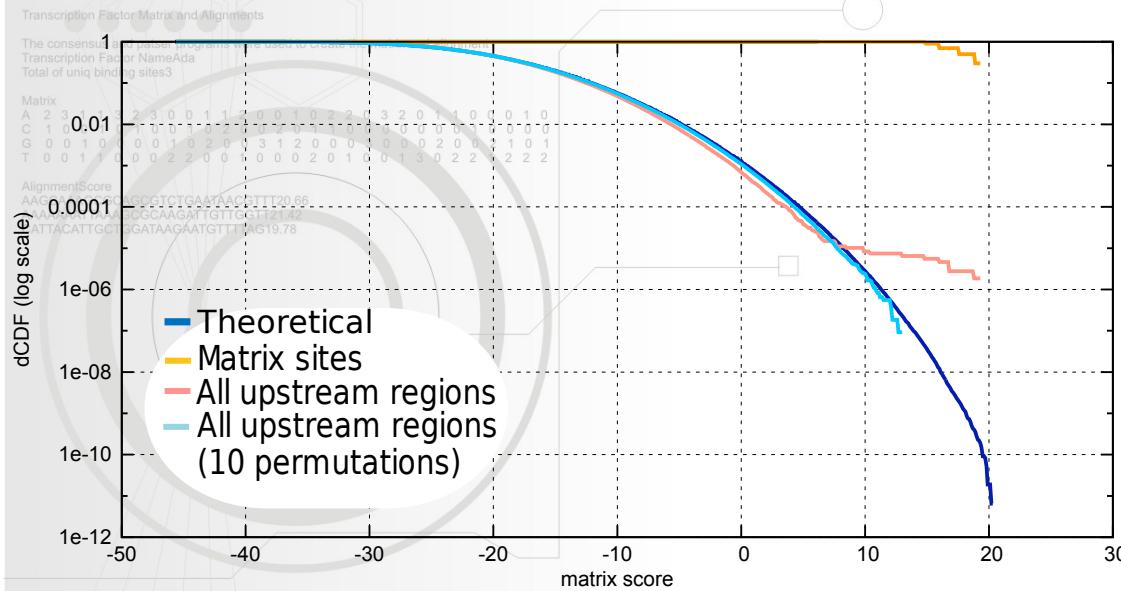
Matrix sites

Release: 6.4 Date: 10-AUG-09

RegulonDB: A Database of Promoter Sequences and Transcriptional Network

Transcription Factor Matrix and Alignments

GAACTAGTTAACTAGTACG      19.3  
GTACTCTTAGCGAGTACA      14.9



## Empirical score distribution in upstream regions

The ‘empirical score distribution in the annotated binding sites’ indicates the sensitivity of the matrix, i.e. its capability to recover binding sites above a given WS threshold.

# Empirical score distribution in the annotated binding sites

## Leave-One-Out (LOO) Test

### Theoretical score distribution

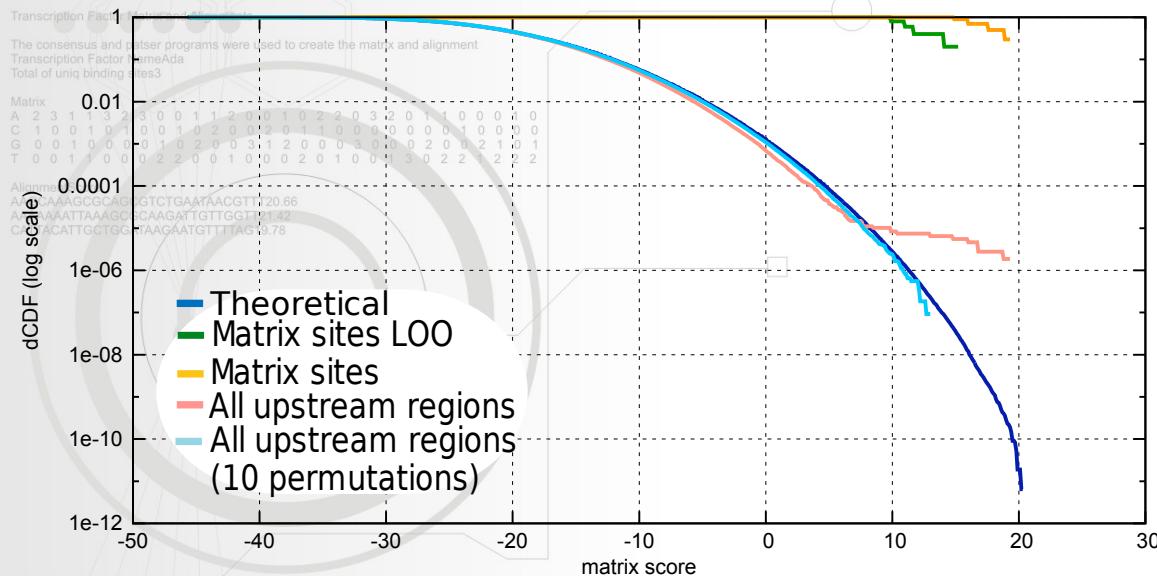
Negative control: Permuted matrix.

Matrix sites

Leave One Out

### Empirical score distribution in upstream regions

**GAACTAGTTAACTAGTACG**    15.1  
**GTACTCTTAGCGAGTACA**    9.8



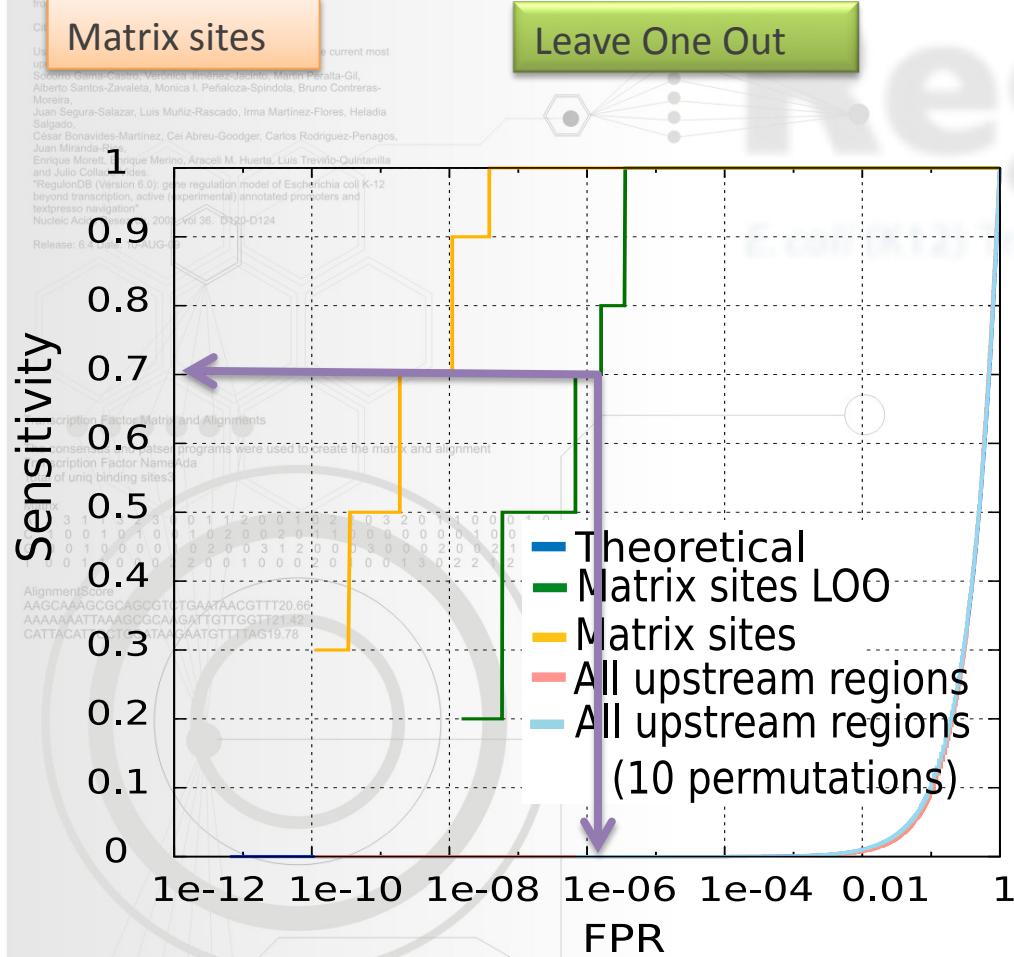
Matrices are rebuilt and annotated sites are scored using a LOO procedure to reduce over-fitting biases when estimating the capability to detect novel sites.

# Receiver Operating Characteristic (ROC) curves

Copies and Copyright-Notice

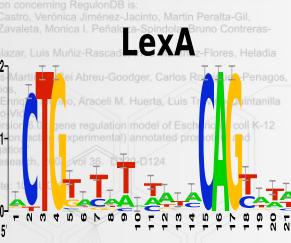
User is not entitled to change or erase data sets of the RegulonDB databases or to eliminate copyright notices from RegulonDB. Furthermore, User is not entitled to expand RegulonDB or to integrate RegulonDB partly or as a whole into other database systems, without prior written consent from CCG.

CCG reserves the right to change or update the current most up-to-date version of the RegulonDB database at any time. The current most up-to-date version of the RegulonDB database is available at <http://regulondb.ccg.unam.mx>.  
 Soledad Gómez-Castro, Verónica Almendro-Jiménez, Martín Peralta-Gil, Alberto Santos-Zavaleta, Mónica I. Pérez-Orive, Bruno Contreras-Moreira, Juan Segura-Salazar, Luis Muñiz-Rascado, Irma Martínez-Flores, Heladio Salgado, Carlos Bonavides-Martínez, Cei Abreu-Goodger, Carlos Rodríguez-Penagos, Juan Miranda-Bonilla, Enrique Moret, Cirio Merino, Ángeles M. Huerta, Luis Treviño-Quiñanilla and Julio Collantes.  
 "RegulonDB (Version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active experimental annotated promoters and transcriptome navigation". Nucleic Acids Res. 2010 Jan; 38: D420-D424.  
 Release: C-2010-01-01-0124

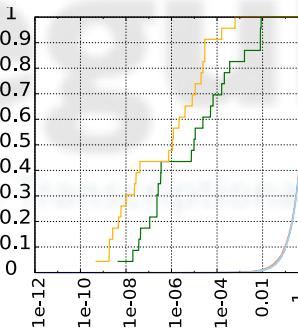
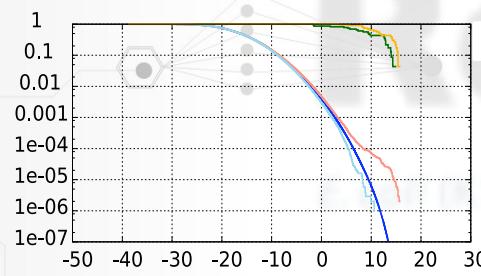


- Receiver Operating Characteristic (ROC) curves' are drawn to indicate the tradeoff between sensitivity and False Positive Rate (FPR). These curves provide a direct way to estimate the expected cost (in terms of false positives) for achieving a desired sensitivity, or, reciprocally, the sensitivity that can be expected for a given FPR.

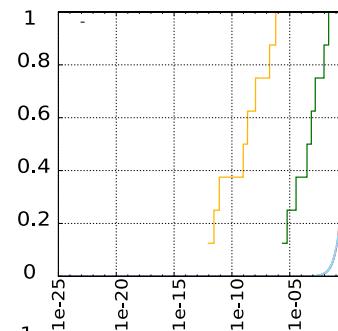
The LexA logo features the word "LexA" in a large, bold, black font. The letter "A" has a vertical bar extending downwards from its top. Below the main title, there is a smaller, semi-transparent watermark-like text that reads: "User is committed to cite properly the use of RegulonDB. The current most update publication concerning RegulonDB is: Socorro Gama-Castro, Verónica Jiménez-Jacinto, Martín Perellá-García, Alberto Santos-Zavala, Monica I. Pérez, and Francisco J. Benítez-Contreras. *J. Segura-Salazar, Luis Muñiz-Rascado, M. Flores, Héledia Salgado, César Bonavides-Martínez, Leil Abreu-Goodger, Carlos Rodríguez-Penagos, and María Jimena Rodríguez. *RegulonDB: a regulatory module of Escherichia coli K-12 genome-wide transcriptional regulation. An experimental annotated database expressed in a systems biology framework*. Nucleic Acids Research 32, Database issue, vol. 32, pp. D1-D124.*



# CysB



- Matrices with information.
  - Low FPR.
  - Detects sites in the genome.
  - LOO ROC is not separated by orders of magnitude from the matrix-sites ROC.



- Matrices with poor information.
  - High FPR
  - Does not detect sites in the genome.
  - LOO ROC is separated by orders of magnitude from the matrix-sites ROC.