



TRANSCRIPTION FACTOR WEIGHT MATRIX EVALUATION

Alejandra Eugenia Medina Rivera

RegulonDB database



ULB



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 f...
database or to eliminate copyright notices from regulon
User is not entitled to expand RegulonDB or to integrate
or as a whole into other databank systems, without prior
from CCG-UNAM.

Citation



A Annotated TrpR binding sites

Site ID

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
aroH
aroH
mtr
mtr
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

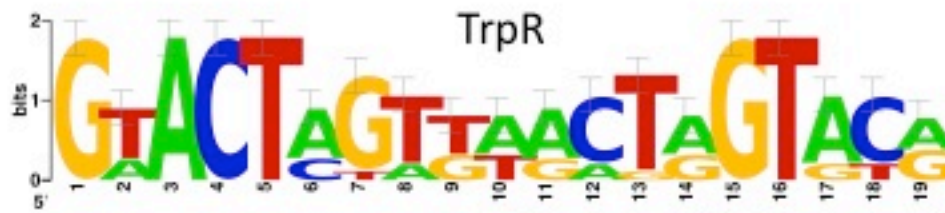
C Consensus

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

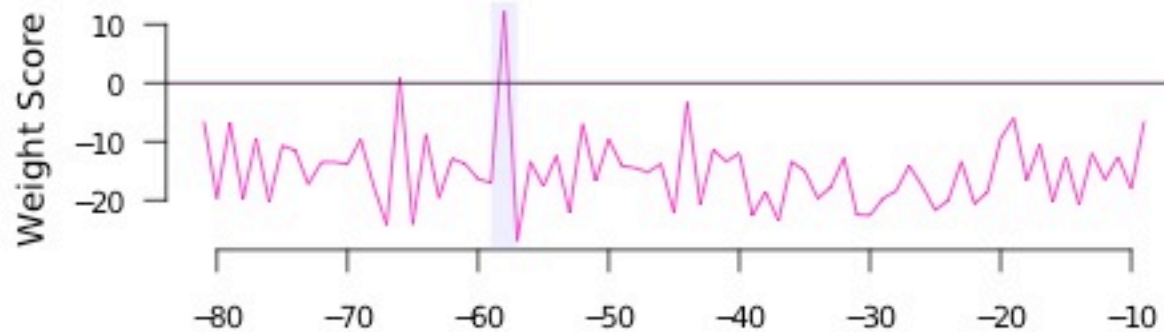
D Sequence logo



Pattern-matching: scanning sequences for putative TFBSs



Search TrpR binding sites



Position in *trpR* upstream sequence (0 corresponds to TSS)

Matrix quality

- 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**?

Evaluation

¹Bailey and Elkan. Systems for Molecular Biology (1994)

²Schneider *et al.* J Mol Biol (1986)

Copies and Copyright-Notice

User is not entitled to change or erase data sets of the RegulonDB database 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 Martínez-García, Roberto Díaz, Alberto Santos-Zavaleta, Monica Delgado, Roberto Martínez-Moreira, Juan Segura-Salazar, Luis Muñoz-Romero, Irma Martínez-Vives, Helena Salgado, César Bonavides-Martínez, Cel Abreu-Goodger, Carlos Rodríguez-Penagos, Juan Miranda-Rios, Enrique Morett, Enrique Merino, Araceli 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 textpresso navigation"
Nucleic Acids Research, 2009, vol 36, D122-D124

Release: 6.4 Date: 10-AUG-09

Transcription Factor Matrix and Alignments

The consensus and patser programs were used to calculate the transcription factor binding sites.
Total of uniq binding sites: 3

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 1
G 0 0 1 0 0 0 0 1 0 2 0 0 3 1 2 0
T 0 0 1 1 0 0 0 2 2 0 0 1 0 0 0 2

AlignmentScore

AAGCAAAGCGACGCTCTGAATAACGTTT20.66
AAAAAATTAAGGCGCAAGATTGTTGGTT21.42
CATTACATGCTGGATAAGATGTTTAG19.78

- Combines theoretical and empirical information to evaluate the quality of matrices, with a series of graphs .

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^{1,2,*}, Cel Abreu-Goodger³, Morgane Thomas-Chollier⁴, Heladia Salgado¹, Julio Collado-Vides¹ and Jacques van Helden^{1,2}

¹Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México. Av. Universidad s/n. Cuernavaca, Col. Chamilpa, Morelos 62210; Mexico, ²Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRé), Université Libre de Bruxelles, Campus Plaine, CP 263. Bld du Triomphe. B-1050 Bruxelles, Belgium, ³EMBL—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK and ⁴Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrass 73. 14195 Berlin, Germany

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

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

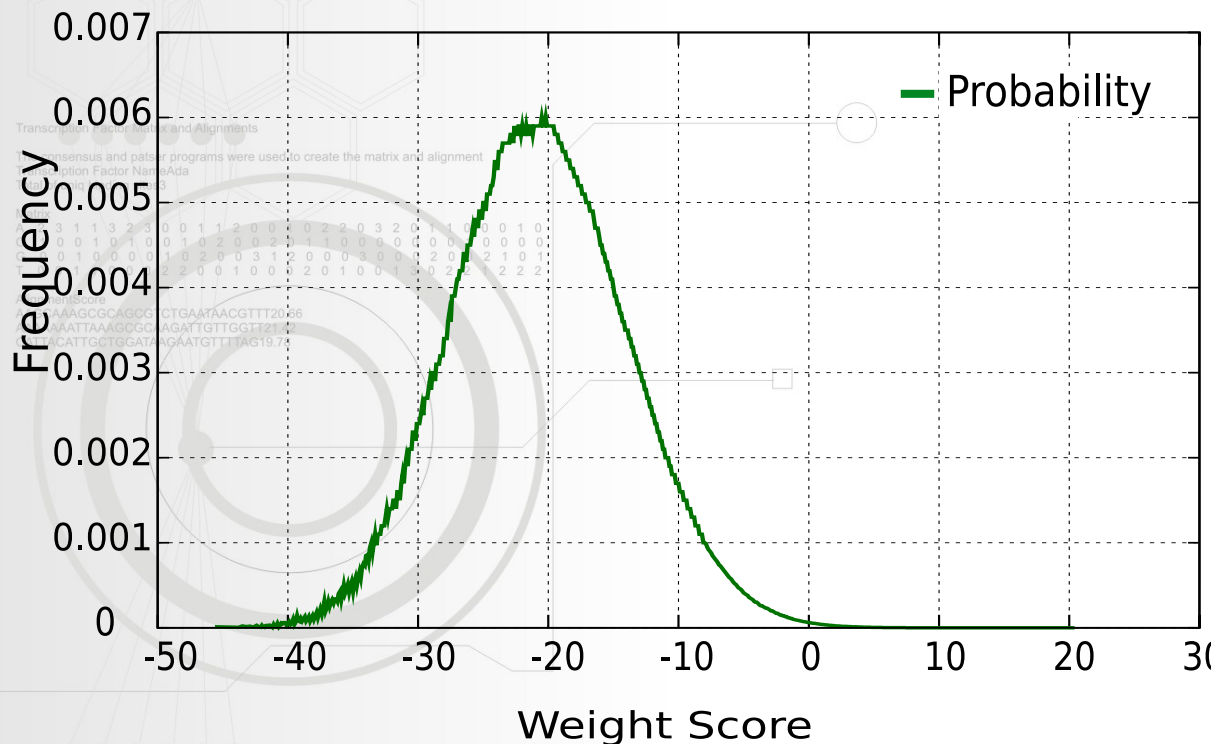
Score distribution: Theoretical distribution

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	-0.07	0.79	-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	-0.04	1.09	-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	-0.07	0.25	1.42	-0.04	0	-0.04	0.41

ATATACGTATCTACTACTG =

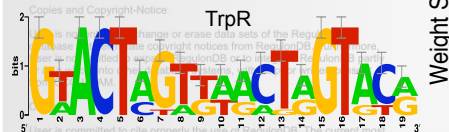
3.25

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

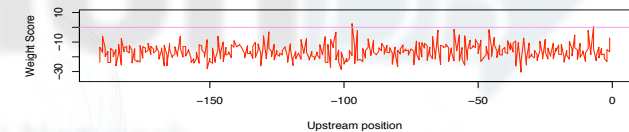
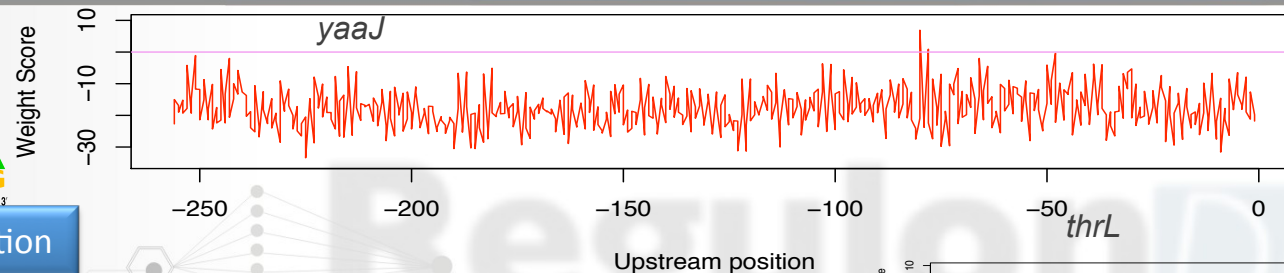
Staden. Comput Appl Biosci (1989)
Extended to higher markov models : *matrix-distrib* (RSAT)



Theoretical score distribution

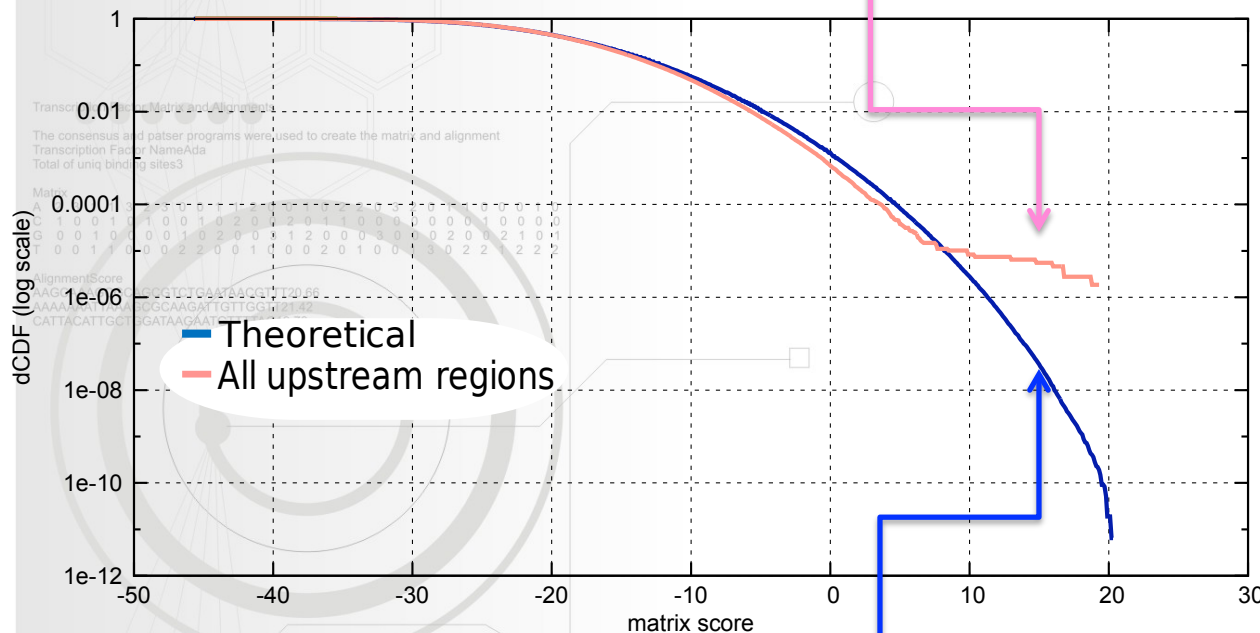
César Bonavides-Martínez, Cei Abreu-Goodger, Carlos Rodríguez-Penagos

Empirical score



Upstream position
... 4404 upstream regions

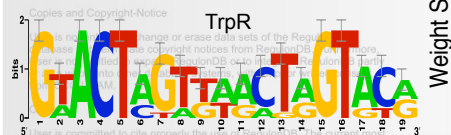
Observed Sites with Score 15 = 7 sites



Expected Sites with Score 15 = 0.04 sites

The 'empirical score distribution in all upstream non-coding sequences' of the organism of interest. These sequences are essentially composed of non-binding sites (the non-coding genomic background), interspersed with a few functional binding sites. The empirical distribution typically fits the theoretical distribution for small WS values (the background), but separates at high WS values, most likely corresponding to functional TF-binding sites.⁹

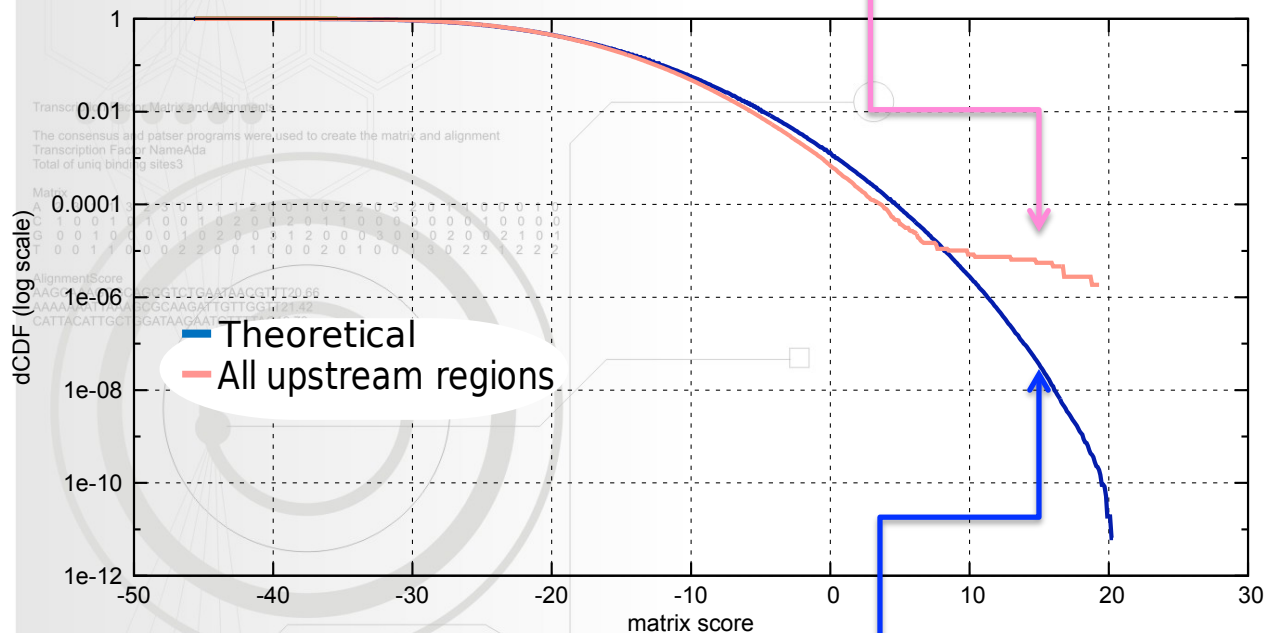
Empirical score distribution in all upstream non-coding sequences



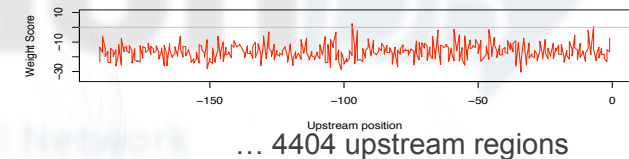
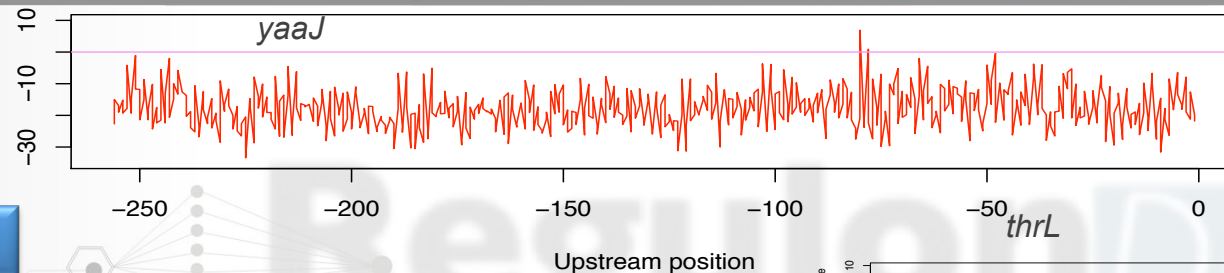
Theoretical score distribution

Empirical score distribution in upstream regions

Observed Sites with Score 15 = 7 sites



Expected Sites with Score 15 = 0.04 sites



... 4404 upstream regions

The 'separation between the right tails of the empirical and theoretical distributions' indicates the capability of the matrix to identify a set of high-scoring putative binding sites in the collection of promoters

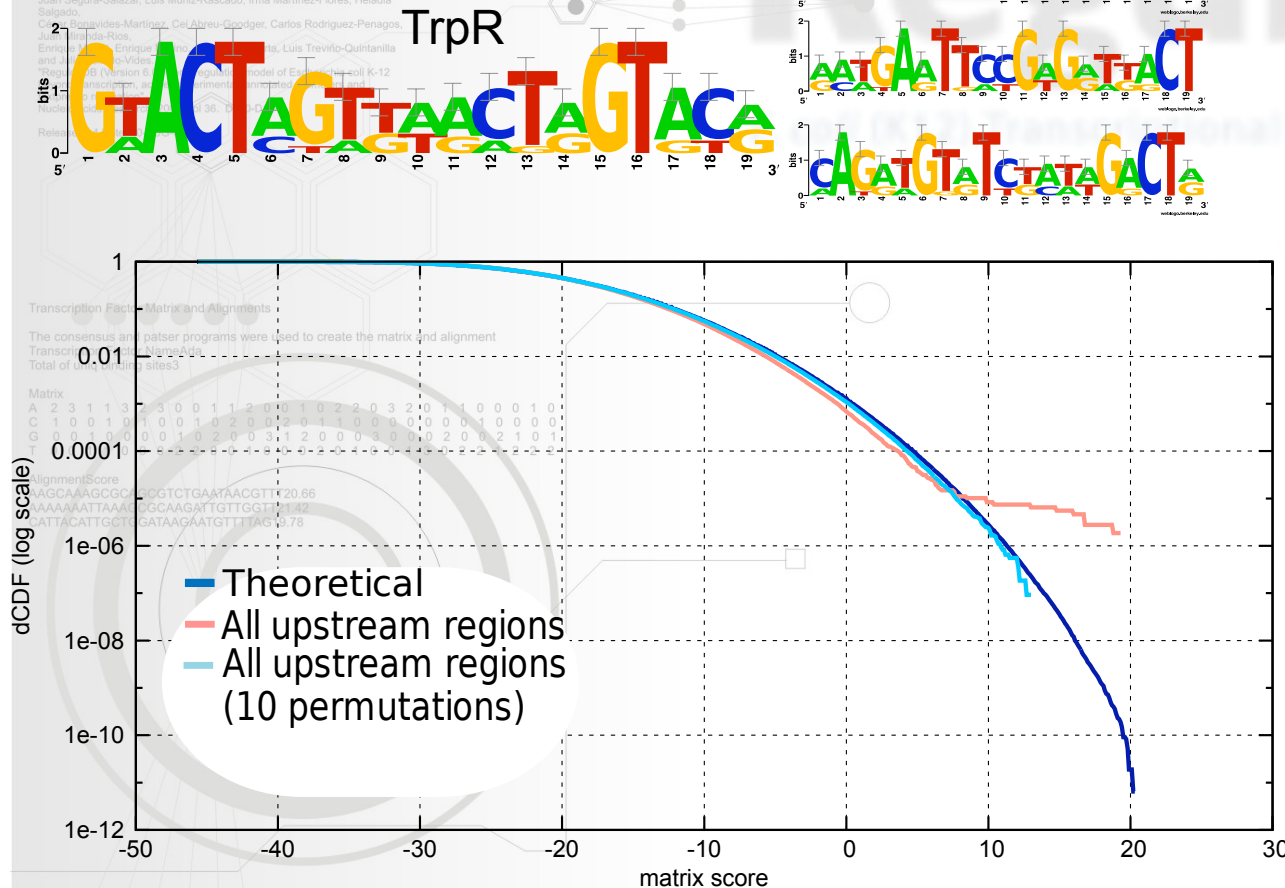
Empirical score distribution in all upstream non-coding sequences

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

Leave-One-Out (LOO) Test

Theoretical score distribution

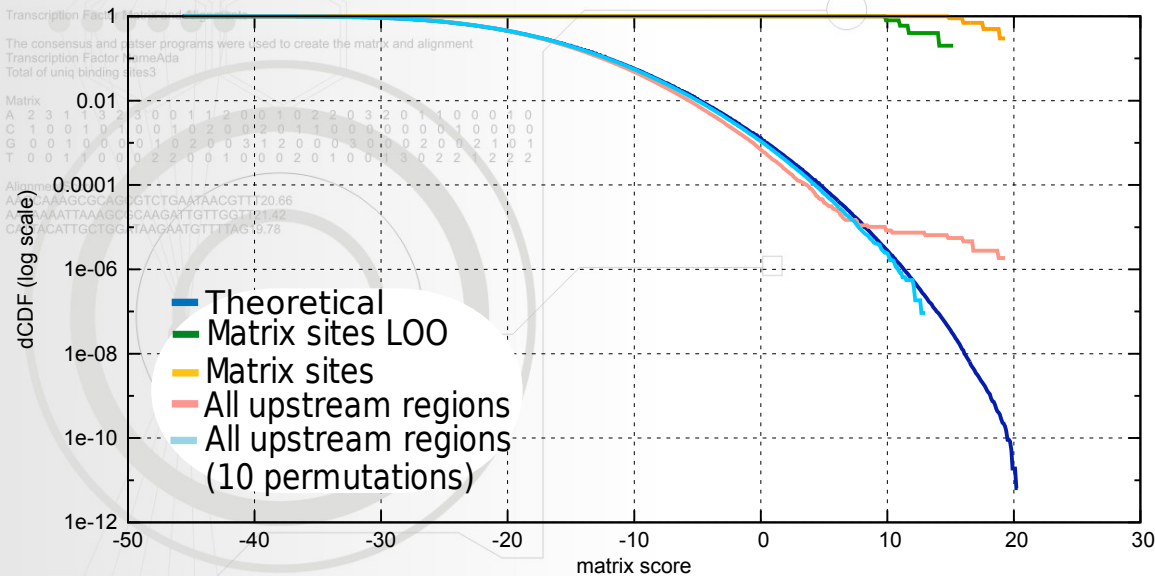
Empirical score distribution in upstream regions

Negative control: Permuted matrix.

Matrix sites

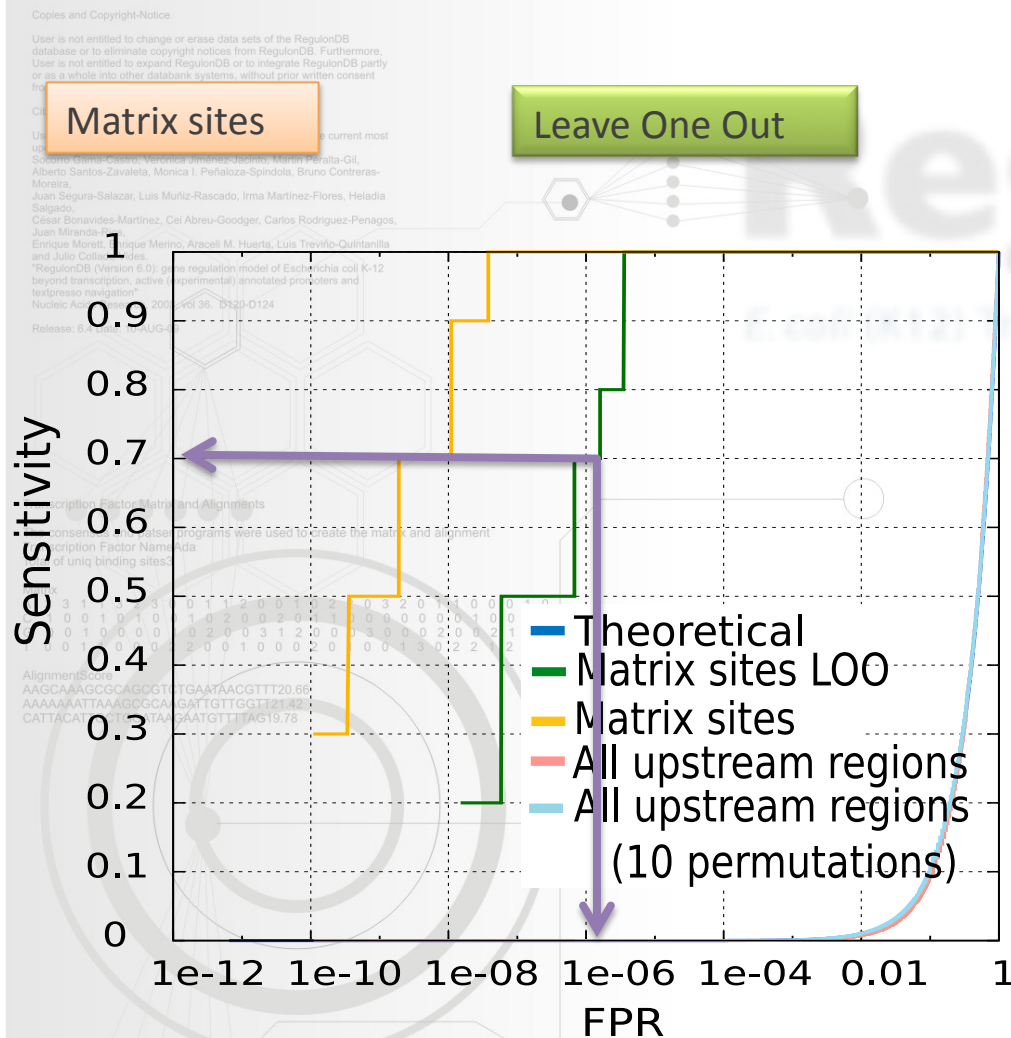
Leave One Out

GAACTAGTTAACTAGTACG 15.1
GTACTCTTTAGCGAGTACA 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



- 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.

matrix-quality in RegulonDB

Evaluation criteria

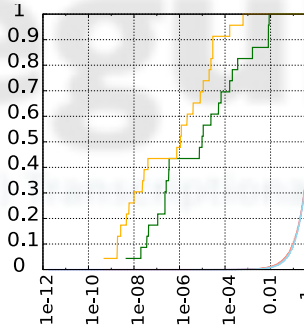
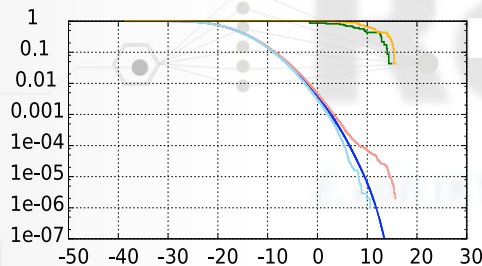
Copies and Copyright-Notice

User is not entitled to change or erase data sets of the RegulonDB database 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, Marín Peralta-Gil, Alberto Santos-Zavaleta, Monica I. Peña, Salvador Contreras-Moreira, Juan Segura-Salazar, Luis Muñoz-Rascón, Flores, Heladia Salgado, César Bonavides-Martínez, Abreu-Goodger, Carlos Ruiz-Peñagos, Juan Miranda-Rosado, Enrique Morett, Enrique Arce, Araceli M. Huerta, Luis Collado-Vides, and Julio Collado-Vides. "RegulonDB: the gene regulation model of Escherichia coli K-12 beyond transcription: experimental, annotated, and predicted regulatory interactions in the genome." Nucleic Acids Research 36(1):1-10 (2008).
Release: 6.4 Data

LexA



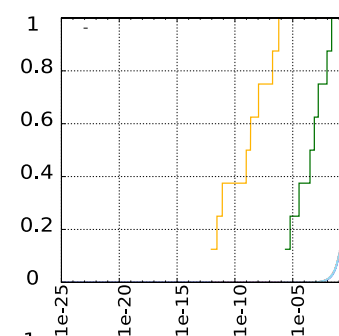
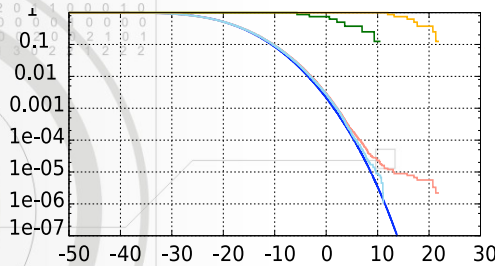
- Matrices with information.
- Low FPR.
- Detects sites in the genome.
- LOO ROC is not separated by orders of magnitude from the matrix-sites ROC.

Transcription Factor Matrix and Alignments

The consensus and patser programs were used to create the matrix and alignment
Transcription Factor NameAda
Total of uniq binding sites3

Matrix
A 2 3 1 1 3 2 3 0 0 1 1 2 0 0 0 1 0 2 2 0 3 2 0 1 1 0 0 0 0 1 0
C 1 0 0 1 0 1 0 0 0 1 1 2 0 0 0 3 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0
G 0 0 1 1 0 0 0 0 1 1 1 2 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
T 0 0 1 1 0 0 0 0 2 2 1 0 0 0 0 2 0 0 1 0 0 0 1
Alignment score
AAGCAAAACGCGACGCTCTGAATAACGTTT20.66
AAAAAAAATAAGGGCAAGATTGTTGGT21.42
CATTACATTCGTCGATAAAGATGTTTAA19.78

CysB



- 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.