DATASET DESCRIPTION

DATASET:

COMPUTATIONAL PREDICTED TRANSCRIPTION FACTOR BINDING SITES FOR 86 TFs IN THE *E. coli* UPSTREAM REGIONS OF THE GENOME. VERSION 2.0

Contact person for this dataset:

Person: RegulonDB team

Email address: regulondb@ccg.unam.mx

Type of dataset:

Computationally predicted transcription factor binding sites (TFBSs).

Reference:

Medina-Rivera et al. Theoretical and empirical quality assessment of transcription factor-binding motifs. Nucleic Acids Research (2011) vol. 39 (3) pp. 808-824

Description:

Computationally predicted transcription factor binding sites (TFBSs) for 86 TFs in upstream regions of the *Escherichia coli* K-12 genome, based on version 7.4 of RegulonDB. A total 8,718 of predicted TFBSs were generated.

Summary:

A total of 86 matrices were used to generate this dataset. One matrix was built for each TF with more than 4 none overlapped annotated binding sites. We built matrices varying parameters as program (MEME, consensus), background model (order cero, order one) and width (±4 pbs of the annotated width). We evaluated different matrices for each TF and selected the one with the best quality (Medina-Rivera et al, 2011) and we were able to obtain a high quality matrix for 50% of the TFs.

In the previous version 71 TFs had enough sites to built a matrix, the 52% of those matrices where of high quality, in the new version for this subset of TFs the percentage has been increased to 60%.

Using the evaluated set of matrices we found 8,718 predicted binding sites, with a specific P-value determined on basis of the results obtained using *matrix-quality*.

Description	**TFBSs RegulonDB**	**TFBSs predictions**

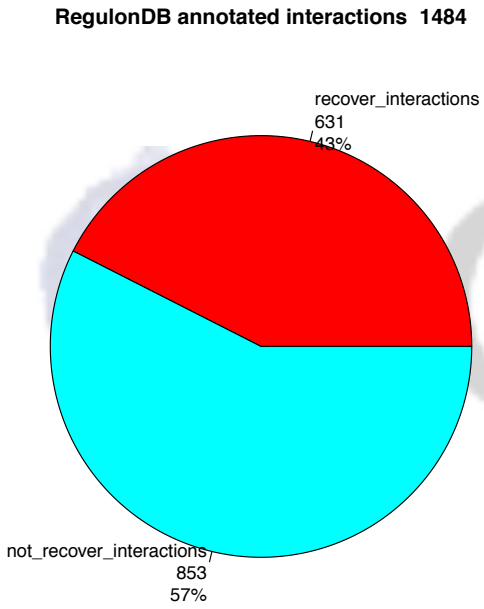
Total of interactions	1484 (100%)	7,415 (100%)
Total of interactions in RegulonDB recovered	631 (43%)	631 (9 %)
Total of interactions not recovered	853 (57%)	6784 (91 %)

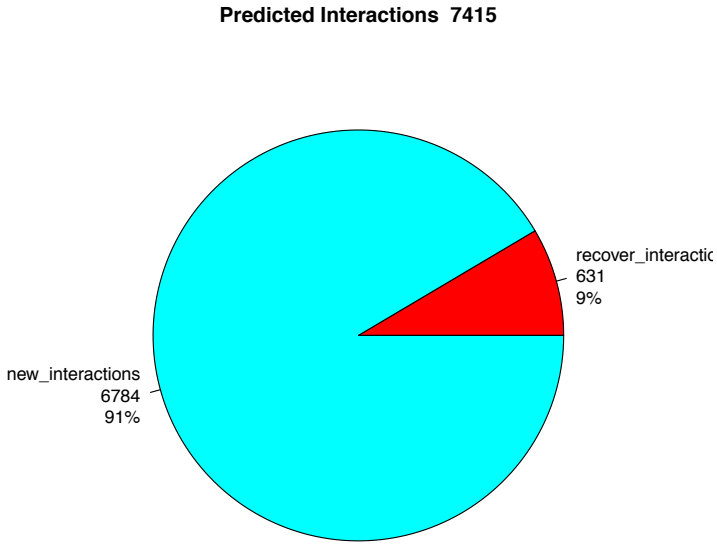
Total of predicted interactions: 7,415

Total of interactions reported in RegulonDB: 1484

Total of interactions in RegulonB that were recovered in the predictions: 631

New predicted interactions: 6784





Methods

Version of programs:

This is version X of the collection of matrices, generated with CONSENSUS version 1E, and MEME version 4.1, and evaluated with *matrix-quality* version 1.141. Predictions for TFBSs were identified with *matrix-scan* version 1.159.

Version of datasets:

Predictions were generated in the regulatory sequences of *Escherichia coli* K12 genome substr MG1655 uid57779 version 2011_08_02.204210. The known TFBSs were obtained from RegulonDB dataset version 7.4.

Protocol

1) The version of the collections of PSSMs is the 2.4, described in the document MatricesRegulonDB_R_v0.2.4.docx

2) Upstream region sequences from *E. coli* K12 are retrieved using the command *retrieve-seq* from the RSAT suite. Upstream sequences are defined as: the sequences upstream the start codon of the gene and up to -400 pbs, if there is an ORF before the -400 bps, the sequence is cut before the ORF overlaps the region.

3) The selected matrix is used to scan upstream region sequences form the *E. coli* K12 genome.

4) Sites are selected on basis of their P-value. The threshold is decided specifically for each matrix, this threshold is reported in the file data/Matrices_manual_selection.txt

    a. make -f makefiles/scan_sequence_sets.mk matrix_scan_all_TFs

5) Analysis of the recovery of the TRN is done using the following programs:

    a. source(R-scripts/Regulon_interactions_evidence.R)

    b. make –f makefiles/scan_sequence_sets.mk interaction_profile_allTFs_allgenes

    c. make -f makefiles/scan_sequence_sets.mk get_regulondb_profile_recovery

    d. make -f makefiles/scan_sequence_sets.mk eval_matrix_scan_recovery

    e. To count the interactions use: source (R-scripts/analysis_of_compare_classes.R)

Prediction of TFBSs

We scanned all upstream regions of every single gene, from -1 to -400 or from -1 to the closest upstream ORF, whatever happens first. This region contains ~80% of currently known TFBSs in RegulonDB-EcoCyc (versions X and Y respectively).

Specificity and sensitivity:

Qualities of matrices used for this propose was evaluated on basis of their sensitivity to recognize known sites and to find them in the genome.

The upper threshold on P-value for sequence scanning was set depending on the evaluation of the quality of the matrix.

Quality of Evidence:

Based on rules in RegulonDB for quality of evidences, all computationally predicted objects are considered WEAK evidence.

Description of the content of the file, column by column

Name of TF	Name of TF motif searched on the sequences.
Lend	Left absolute coordinate of the TF site
Rend	Right absolute coordinate of the TF site

Strand	Strand where the predicted TFBs is located. "F" for Forward, and "R" for Reverse.
Gene name	Name of the associated downstream gene
Gene_strand	Strand of the gene. "F" for Forward, and "R" for Reverse.
Left_relative	Left relative position of the site to the start of the gene
Right_relative	Right relative position of the site to the start of the gene
Site Sequence	Sequence of the predicted binding site
Score	Score of the predicted binding site.
Range of score	Minimal and maximal scores found by the matrix
P-value	Probably of being mistaken while taking the site a real

Citation:

Dataset provided and maintained by RegulonDB (PUBMED: #18158297) from the original source published in:

Medina-Rivera et al. Theoretical and empirical quality assessment of transcription factor-binding motifs. Nucleic Acids Research (2011) vol. 39 (3) pp. 808-824

Licensing:

See the license of RegulonDB in:

http://regulondb.ccg.unam.mx/LicenseRegulonDBWithoutSing.jsp