Metadata for Dataset

General information

Title	Position specific scoring matrices (PSSMs) for TFs, version 4.0
Version	4.0
ContactPersonEma	regulondb@ccg.unam.mx
Creation date	2019-03-29
Format	regulondb@ccg.unam.mx
Creation date	plain text
Language	English
Rights	Dataset provided and maintained by RegulonDB (PUBMED: #21051347) 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 (PubMed: 20923783)
Citation	CC BY-NC-SA 4.0: https://creativecommons.org/licenses/by-nc-sa/4.0

Related Reference or Resources

RelatedIdentifier	http://regulondb.ccg.unam.mx/menu/download/datasets/files/BindingSiteSet.txt
RelatedIdentifierTy	URL
RelationType	IsDerivedFrom
RelatedIdentifier	http://embnet.ccg.unam.mx/rsat/
RelatedIdentifierTy	URL
RelationType	IsPartOf [IsResultFrom]
RelatedIdentifier	20923783
RelatedIdentifierTy	PMID
RelationType	IsPartOf

Authors and co-authors

Creator	Medina-Rivera, Alejandra
Affiliation	International Laboratory for Human Genome Research (UNAM)
Creator	Gonzalez, Cristian
Affiliation	Center for Genome Sciences (UNAM)
Contributor	RegulonDB
ContributorType	ResearchGroup

Dataset Details

A. Description

The PSSMs version 4.0 are built using the annotated binding sites of TFs from RegulonDB version 10.5.2. A matrix is built for all TFs with four or more annotated sites. Each matrix is shown with theoretical and empirical quality assessment.

B. Data Summary

Description	Total
Total number of Matrices	93
Total matrices with good or acceptable quality	65
Total matrices with low quality	28
Total matrices with a previous matrix	80
Total matrices with good or acceptable quality with a previous matrix	50
Improved Matrices: low quality to acceptable	13
Improved Matrices: low quality to good quality	5
Improved Matrices: acceptable quality to good quality	11
TFs with previous matrices and high quality	19

C. Method

C.1. Programs

1. The MEME version 5.0.4 is executed to generate different size of matrices.
2. The RSAT/matrix_quality version 1.0 evaluates the quality of the matrices.
3. The best evaluated matrix is choosen for each TF.

C.2. Datasets, databases, files

1. Genome Sequence
   1.1. Genome Sequence Identifier: NC_000913.3
   1.2. E.coli's Genome Version: version 3 (NC_000913.3)
2. RegulonDB dataset
   2.1. BindingSiteSet.txt version 10.5.2

C.3. Protocol

1. TFBSs are retrieved from RegulonDB web page -BindingSiteSet.txt.
2. A script is execute to run the MEME program to built the matrices for all the TFs.
   2a. An approximate of 30 alternative PSSMs are built for each TF.
3. The matrix-quality program (RSAT) is run to evaluate the quality of each matrix.
4. The compare-qualities program (RSAT) is executed to choose the best matrix.

C.4. Specificity and sensitivity

"For an unbiased estimate of sensitivity, we would ideally need two separate collections of sites: one for building the PSSM, another for testing it. Unfortunately, for most TFs, very few binding sites are known. In order to ensure an independent assessment while minimizing the loss of information, the program 'matrix-quality' performs a LOO validation, iteratively discarding one annotated site, re-building the matrix, and scoring the left-out site with the new matrix. The program also discards multiple copies of identical sites, if those are not from independent sources, which would otherwise induce the same kind of bias. RegulonDB contains 10 TrpR sites, with only five remaining after redundancy filtering. Not surprisingly, when applying the redundancy filter and the LOO procedure these sites have lower scores ranging from 9.62 to 15.78. The LOO score distribution thus corrects obvious biases in the estimation of the matrix sensitivity, and the difference with the matrix sites distribution indicates the level of over-fitting to the training sites." (Medina-Rivera A, 2010)

D. Dataset format

D.1 Text File (.txt)

Columns Description:

1. Transcription Factor Identifier
2. Transcription Factor Name
3. Total of binding sites
4. PSSM size
5. Alignment
6. PSSM

D.2 Consensus File (.cons)

Description:
; Transcription Factor ID, TFName: Transcription Factor Name, Total of TFBSs: Total of binding sites, PSSM size: Size of the TFBS Position Scoring Site Matriz
//