DATASET DESCRIPTION

DATASET:

Promoter_from_454_Dataset.txt. Experimental determination of Transcription Star Sites (TSSs) with High Throughput Pyrosequencing Strategy (HTPS) and computational promoter prediction.

Contact person for this dataset:

Person: RegulonDB team
Email address: regulondb@ccg.unam.mx

Type of dataset:

Experimental TSS mapping and computational promoter prediction

Reference:

Alfredo Mendoza, Leticia Olvera, Maricela Olvera, Ricardo Grande, Veronica Jiménez-Jacinto, Blanca Taboada, Leticia Vega, Katy Juárez, Heladia Salgado, Araceli Huerta, Julio Collado-Vides and Enrique Morett. (2009). Genome-wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in E. coli . PLoS ONE. 4(10):e7526.

Description:

This file contains:

High Throughput Pyrosequencing Strategy (HTPS) Data, with and without previously know TSSs. For many, promoter region was predicted in this work.

Summary:

This file has a collection of more than 1491 transcription start sites (TSSs) that have been experimentally determined with high precision in Escherichia coli using an unbiased High Throughput Pyrosequencing Strategy (HTPS). From this collection, about 148 corresponded to previously reported TSSs, which helped us to benchmark both our methodologies and the accuracy of the previous mapping experiments. The other ca 1300 TSSs mapped belong to about 900 different genes, many of them with no assigned function.

Methods

Version of programs:

We use blast version BLASTN 2.2.18

For the σ^{38} -10 element a new matrix was constructed with the WCONSENSUS program (version 6d), utilizing the nucleotide sequences of 71 promoters of E. coli annotated in RegulonDB. For this purpose, with a threshold identical to the one used for the –10 element of σ^{70}, we recovered 65% of the promoters of the training set.

With the PATSER program (version 3d) which searches for patterns in a database, and the sequences utilized in this work for the construction of the PWM, the following thresholds were defined: -0.5 SD for the -10 element, and -1 SD for the –35 element. The threshold for the -10 element was chosen because with it, 60% of the promoters used for the training set were recovered. Because the -35 element is less conserved than the –10 element, a lower threshold was selected to avoid false positives. With this value, 40% of the promoters of the initial training set were recovered

Version of datasets:

RegulonDB Release: 6.4 Date: 10-AUG-09

Data Set version 2.0

Protocol or algorithm

With data from Pyrosequencing, we use blast to determined the position and with this, the coverage, from genome. Next develop tools of visualization (GenoSeqGraph), develop some perl script for cut upstream sequence and search new motives, box -10 and box -35, search previously known promoter information and small RNAs, terminators and biologic object. Clear overlapping and make groups.

Quality of Evidence:

Strong and weak evidences

Description of the content of the file, column by column

Column	Name	Description
1	Subset	High Throughput Pyrosequencing Strategy (HTPS) Data, divided in 4 sets: Subset I: With previously known TSS And, With promoter region predicted in this work Subset II: Without previously known TSS and With promoter region predicted in this work Subset III: With previously known TSS and Without promoter region predicted in this work Subset IV: Without previously known TSS and Without promoter region predicted in this work
2	Gene_Name	Gene name from regulonDB
3	Relative_Position_toATG	Relative position to ATG gen
4	Absolute_Position	Genome map position of Transcription Start Site (+1)
5	Frecuency	Frecuency: number of sequence in the same position
6	TSS_Strand	DNA strand sequence
7	Seq_ID	Sequence identifier assigned by Roche 454 GS Sequencer or assigned by authors
8	Genomic Region	Genomic region where is located the TSS with respect to the nearest gene
9	Gene_Position left	Gene left end position in the genome
10	Gene_Position right	Gene rigth end position in the genome
11	Gene_Strand	DNA strand where the gene is coded
12	B_Name	Blattner number (bnumber) of the gene
13	Sequence	Sequence of the promoter region (+1 upper case)
14	Evidence	Type evidence

Citation:

Dataset provided and maintained by RegulonDB (PUBMED: #18158297 http://www.ncbi.nlm.nih.gov/pubmed/18158297?dopt=Abstract) from the original source published in*: Socorro Gama-Castro, Heladia Salgado, Martin Peralta-Gil, Alberto Santos-Zavaleta, Luis Muñiz-Rascado, Hilda Solano-Lira, Verónica Jimenez-Jacinto, Verena Weiss, Jair S. García-Sotelo, Alejandra López-Fuentes, Liliana Porrón-Sotelo, Shirley Alquicira-Hernández, Alejandra Medina-Rivera, Irma Martínez-Flores, Kevin Alquicira-Hernández, César Bonavides-Martínez, Juan Miranda-Ríos, Araceli M. Huerta, Alfredo Mendoza-Vargas, Leonardo Collado-Torres, Blanca Taboada, Leticia Vega-Alvarado, Maricela Olvera, Leticia Olvera, Ricardo Grande, Enrique Morett and Julio Collado-Vides, (2010) RegulonDB version 7.0: Transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor units).*

Licensing:

See the license of RegulonDB in:

http://regulondb.ccg.unam.mx/download/LicenseRegulonDBRegistration.jsp