User´s guide RhizoBindingSites and RhizoBindingSites v2.0

1.- RhizoBindigSites and RhizoBindigSites v2.0 version

2.- Construction of the databases

3.- Mission

4.- Selection of RhizoBindingSites or RhizoBindingSites v2.0 version

5.- Synonyms converter

6.- Criteria for selecting a motif

7.- Motif Information window

8.- Motif map application searches the motif in the upstream region of the orthologs from the Rhizobiales

9.- Motif logo application

10.- Prediction of Regulons

11.- Gene Information window

12.- Matrices window

13.- Example of how to use the matrix-scan data to search who potentially regulates a gene which is not a "TF-gene"

14.- Example of how to use the matrix-scan data to search how the transcriptional regulators of an expected regulon may be interacting

15.- Matrix-clustering

16.- A Brief guide to analyze a genome in the RSAT web site http://embnet.ccg.unam.mx/rsat/

1.- RhizoBindigSites and RhizoBindigSites v2.0 version :
RhizoBindingSites and the v2.0 version databases contain information with a potential role in the transcriptional regulation from nine representative species of the taxon Rhizobiales.

2.- Construction of the databases: RhizoBindingSites database was constructed with matrices deduced from ortholog genes from each gene per genome. In contrast, RhizoBindigSites v2.0 was constructed with matrices by using the short nucleotide sequences from the motif sites for each gene per genome from the matrix-scan output data of RhizoBindingSites. Highlighting, matrices of the RhizoBindingSites v2.0 were deduced from the respective genome sequences.

RhizoBindingSites and RhizoBindingSites v2.0 are revised databases, because all data included is from matrices able to find a motif in their coding string on its regulatory region. As well as, data from matrix-scan deposited in the "Motif Information" window (see below) are with motifs found only on the string codifying the gene from target-genes.

3.- Mission: To provide information with a potential role in transcriptional regulation of genes at three ranges of p-values, from 1.0e-4 to 9.9e-4 (low stringency data), from 1.0e-5 to 9.9e-5 (medium stringency data) and from 1.0e-6 to lower p-values (high stringency data).

4.- Selection of RhizoBindingSites or RhizoBindingSites v2.0 version.
Access to the v2.0 version is by clicking on the legend "Switch to v2" on the main page of database RhizoBindingSites. The functionality of RhizoBindingSites v2.0 windows and the application "Prediction of Transcriptional regulatory networks" is equal to the RhizoBindingSites, se below.

5.- Synonyms converter. For cases in which the User´s Identifiers do not coincide with the ones used in these databases, a list of identifiers synonyms for each of the genomes was included with an application to convert their locus tags to the locus tags used in these databases. Paste their list on the window and run the program, you may copy your equivalent list of Locus tags formats: ***Rhizobium etli*** **CFN42** (RHE_RS), ***Rhizobium etli*** **Mim1** (REMIM1_RS) , ***R. leguminosarum*** **biovar** ***viciae*** **3841** (RL_RS), ***Bradyrhizobium diazoefficiens*** **USDA 110** (AAV28_RS), ***Sinorhizobium fredii*** **NGR234** (NGR_c), ***Sinorhizobium meliloti*** **1021** (SMc), ***Bradyrhizobium sp*** **BTAi1** (BBTA_RS), **Azorhizobium caulinodans ORS 571** (AZC_RS), ***Mesorhizobium japonicum*** **MAFF303099 (**MAFF_RS).

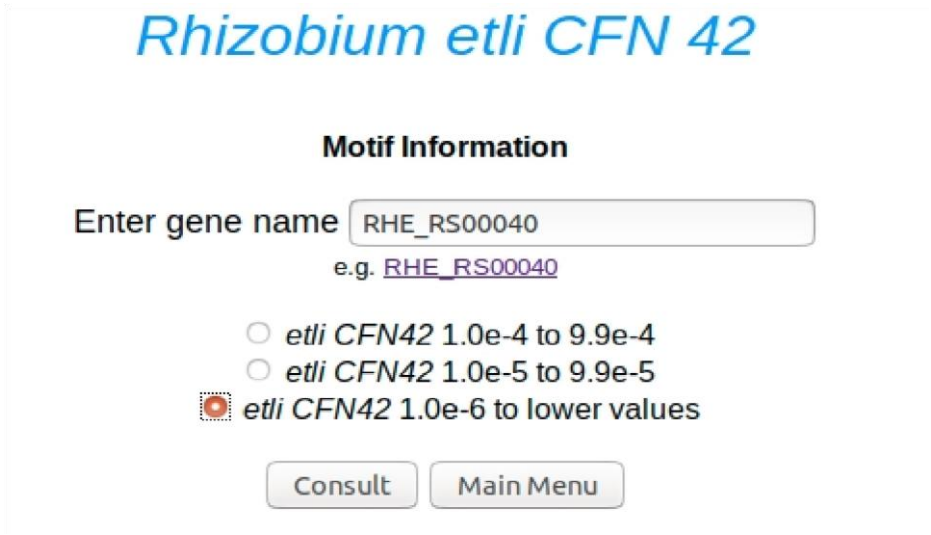
6.- Criteria for selecting a motif: Better selection of a motif should involve many parameters provided in the RhizoBindingSites database such as: motif with the lowest p-value (medium or high stringency data), motif sequence (avoid as possible repeated nucleotide repeated in the motif), conservation of the motif in orthologous genes in the Rhizobiales taxon, coherence of the function of the gene query with genes sharing the motif, look for the query gene has a vicinity with a transcriptional regulator. Additionally, it is advisable to search for information in the scientific literature on the expression in the same physiological condition of both, the transcriptional regulator and the gene-target.

7.- Motif Information window: Provides a table with a locus tag identifier of genes sharing query motifs (e-regulon), gene strand location, matrix identifier, strand location of the matrix, start/end

position of the motif in the gene promoter, nucleotide sequence of the motif, weight, p-value and significance of the site.

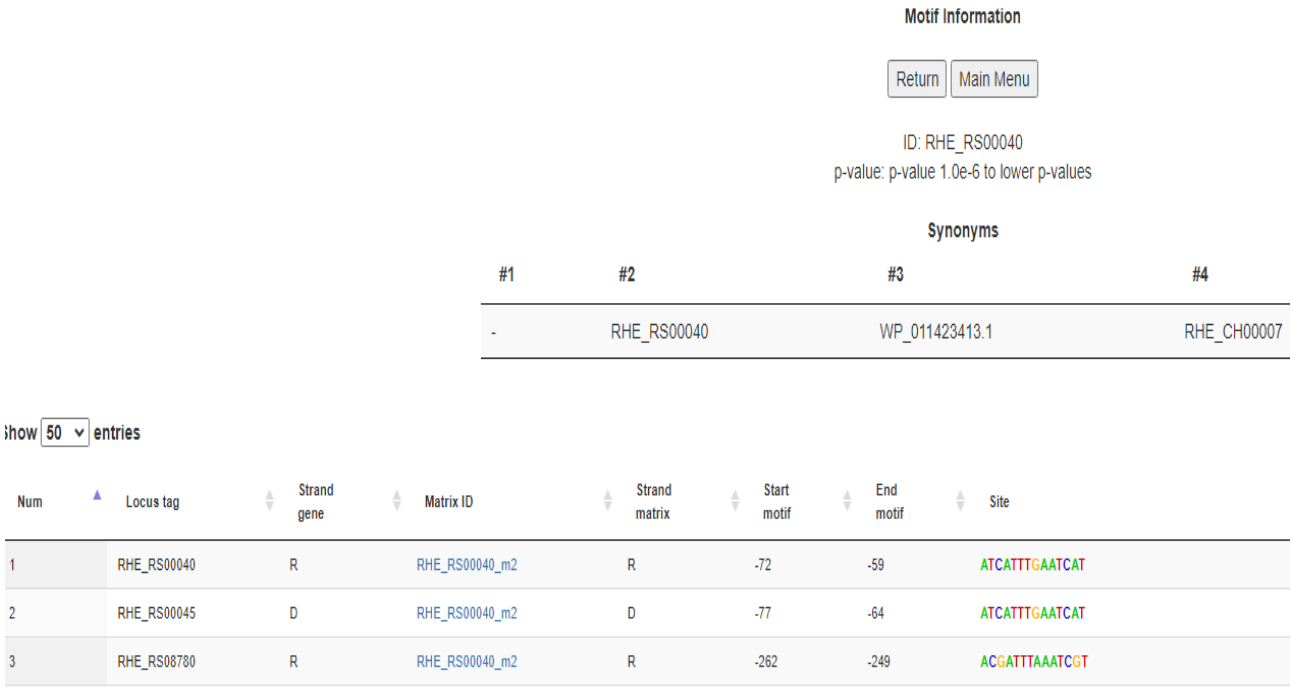
Select a genome.

Enter an NCBI gene identifier in a genome of interest, locus tag, protein ID or gene name or click on the suggested ID.



Select a p-value and click on the "Consult" button. If the gene had no matrices, a legend appears asking if the user wants to look for the presence of motifs in the query promoter gene. The result is provided in a table separated by tabs.

Motif information Table.



Motif Information

Return Main Menu

ID: RHE_RS00040
p-value: p-value 1.0e-6 to lower p-values

Synonyms

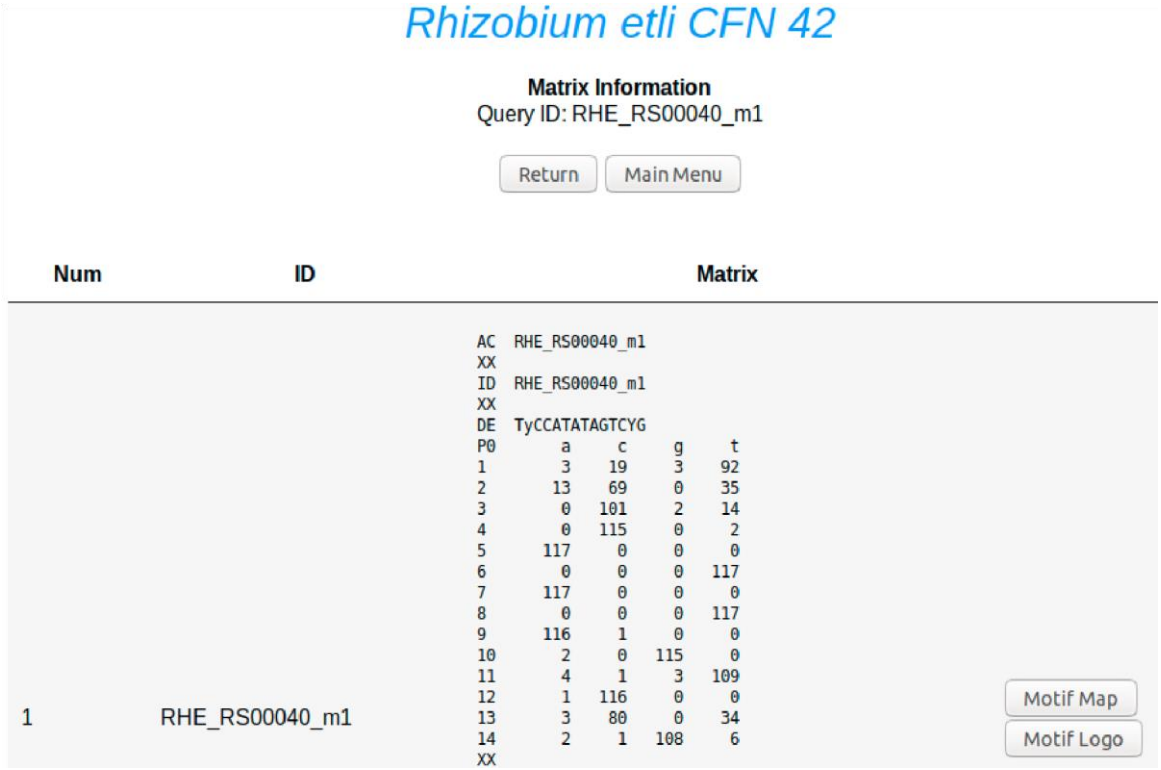
#1	#2	#3	#4
-	RHE_RS00040	WP_011423413.1	RHE_CH00007

Show 50 entries

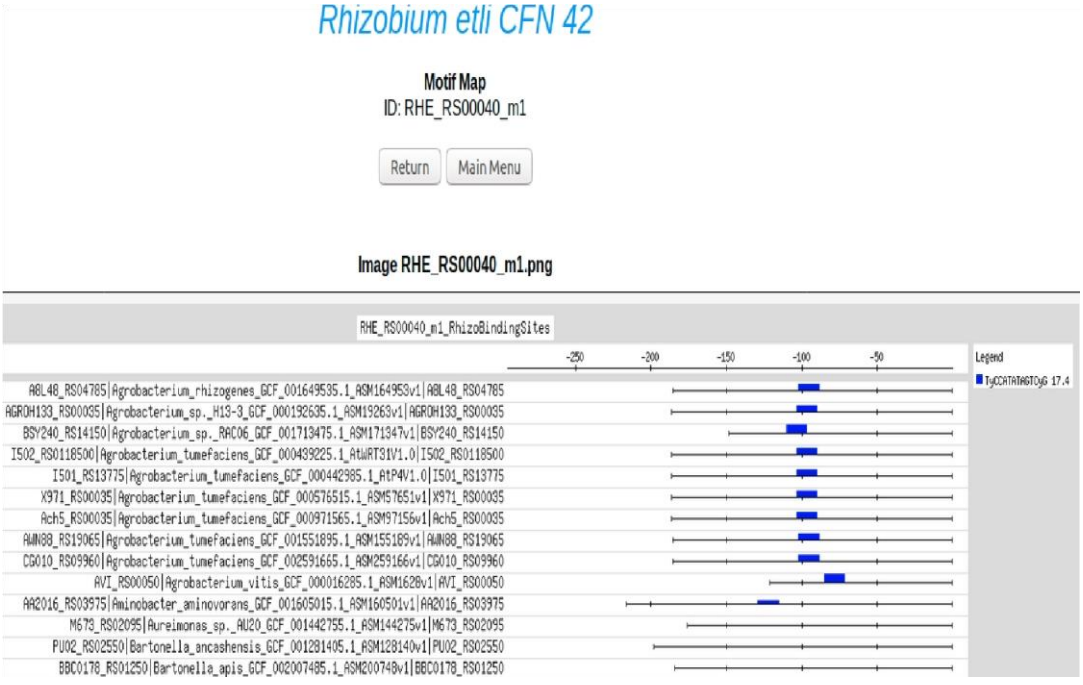
Num	Locus tag	Strand gene	Matrix ID	Strand matrix	Start motif	End motif	Site
1	RHE_RS00040	R	RHE_RS00040_m2	R	-72	-59	ATCATTTGAATCAT
2	RHE_RS00045	D	RHE_RS00040_m2	D	-77	-64	ATCATTTGAATCAT
3	RHE_RS08780	R	RHE_RS00040_m2	R	-262	-249	ACGATTTAAATCGT

8.- Motif map application searches the motif in the upstream region of the orthologs from the query gene in the Rhizobiales taxon. This application is accessed by clicking on a matrix name in the matrix ID column of the table from window “Motif Information”.

The matrix information.



Click on the button “Motif Map” and the conservation of the motif in the Rhizobiales is displayed.



9.- Motif logo application: Displays the Motif logo of the selected matrix on the right and reverse strands of a selected motif by clicking on the “Motif Logo” button.

Rhizobium etli CFN 42

Motif Logo
ID: RHE_RS00040_m1

Return | Main Menu

Image RHE_RS00040_m1_logo.png

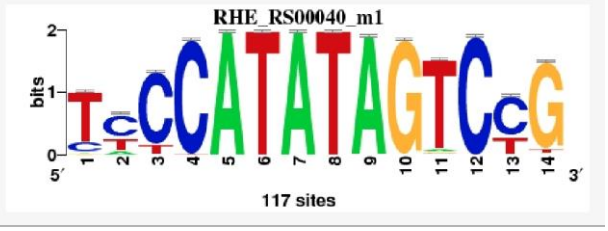
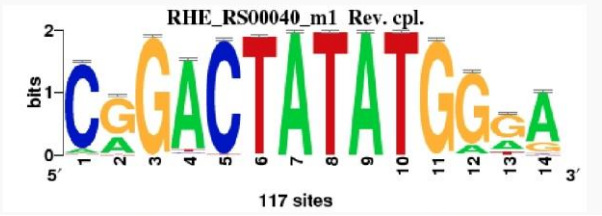
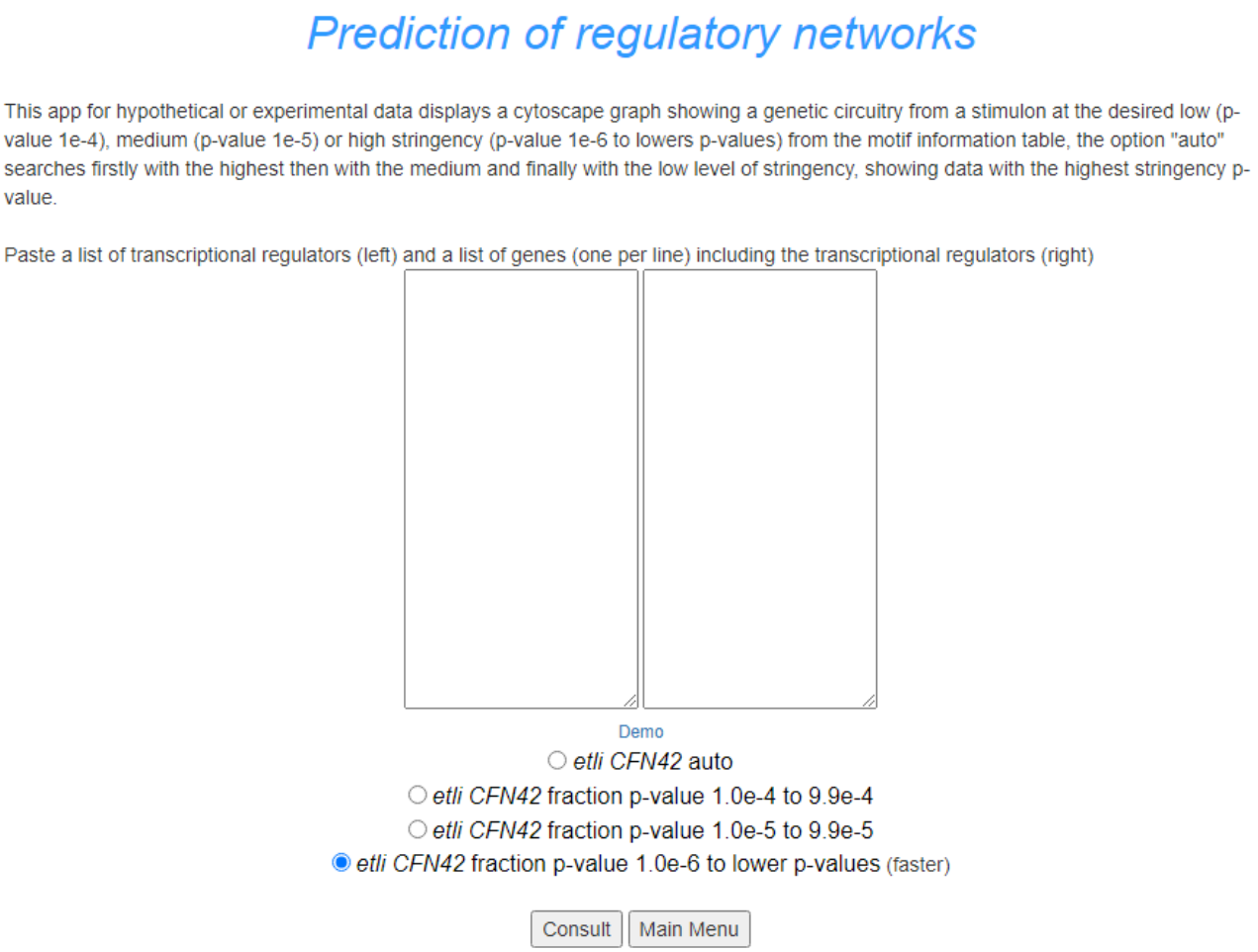


Image RHE_RS00040_m1_logo_rc.png



10.- Prediction of Regulons or transcriptional regulatory networks. This application is in the motif information window. It is for prediction of regulons from experimental or predicted data by writing or pasting a list of locus tags of transcriptional regulators into the left box, and a list of locus tags from the complete list of target proteins, including the transcriptional regulators in the right box. There is a demo in the Genome section of *Rhizobium etli* CFN42. The application for each transcriptional regulator (from the left box) will search for common genes in the data from motif information with the entered list, from the right box, which also contains the transcriptional regulators of the left box.



Select the strain *Rhizobium etli* CFN42. Click on the “demo” link to enter a demo list. Select the “auto” option. Alternatively, you may select a p-value for the search.

Prediction of regulatory networks

This app for hypothetical or experimental data displays a cytoscape graph showing a genetic circuitry from a stimulon at the desired low (p-value 1e-4), medium (p-value 1e-5) or high stringency (p-value 1e-6 to lowers p-values) from the motif information table, the option "auto" searches firstly with the highest then with the medium and finally with the low level of stringency, showing data with the highest stringency p-value.

Paste a list of transcriptional regulators (left) and a list of genes (one per line) including the transcriptional regulators (right)

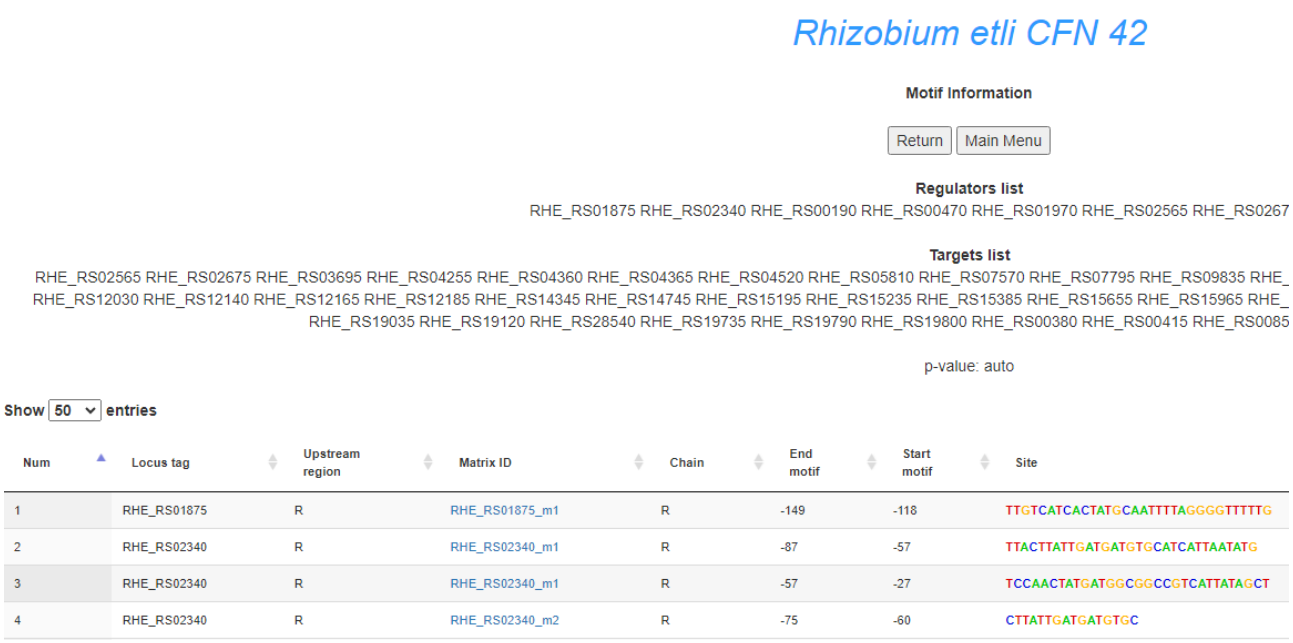
RHE_RS01875 RHE_RS02340 RHE_RS00190 RHE_RS00470 RHE_RS01970 RHE_RS02565 RHE_RS02675 RHE_RS03695	RHE_RS02565 RHE_RS02675 RHE_RS03695 RHE_RS04255 RHE_RS04360 RHE_RS04365 RHE_RS04520 RHE_RS05810 RHE_RS07570 RHE_RS07795 RHE_RS09835 RHE_RS10645 RHE_RS11215 RHE_RS11220 RHE_RS11430

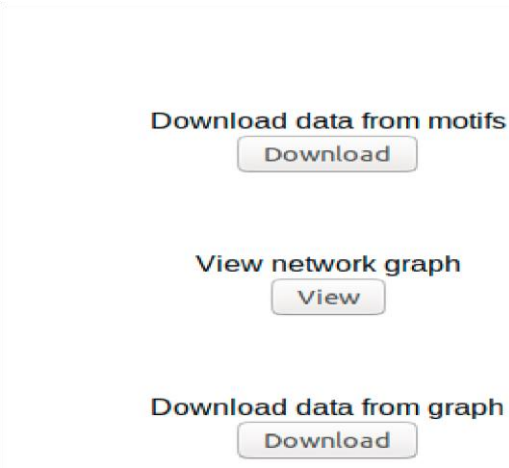
Demo

- (●) *etli CFN42* auto
- ( ) *etli CFN42* fraction p-value 1.0e-4 to 9.9e-4
- ( ) *etli CFN42* fraction p-value 1.0e-5 to 9.9e-5
- ( ) *etli CFN42* fraction p-value 1.0e-6 to lower p-values (faster)

Consult | Main Menu

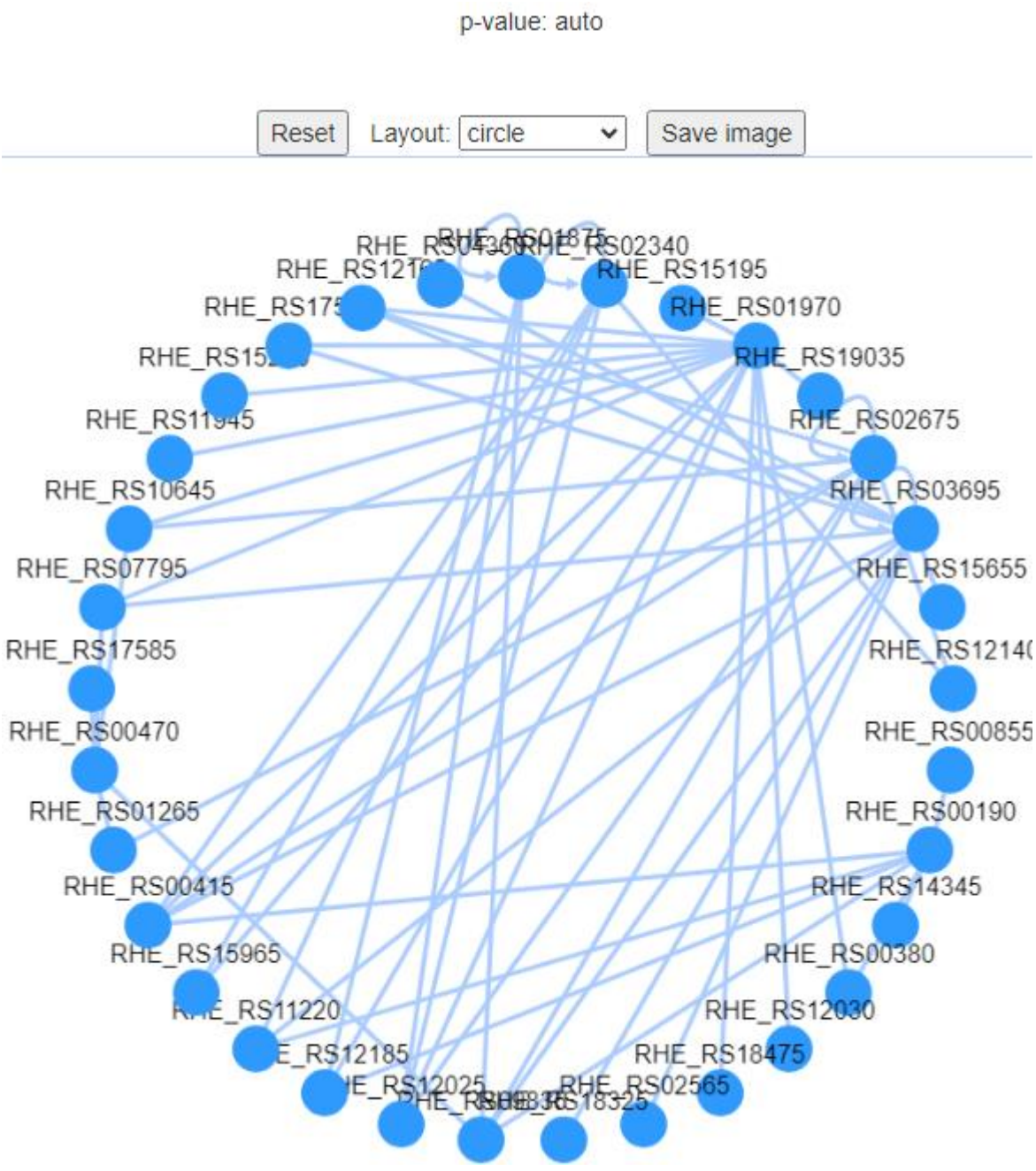
click on the "Consult" button.
Motif information data from the regulon.





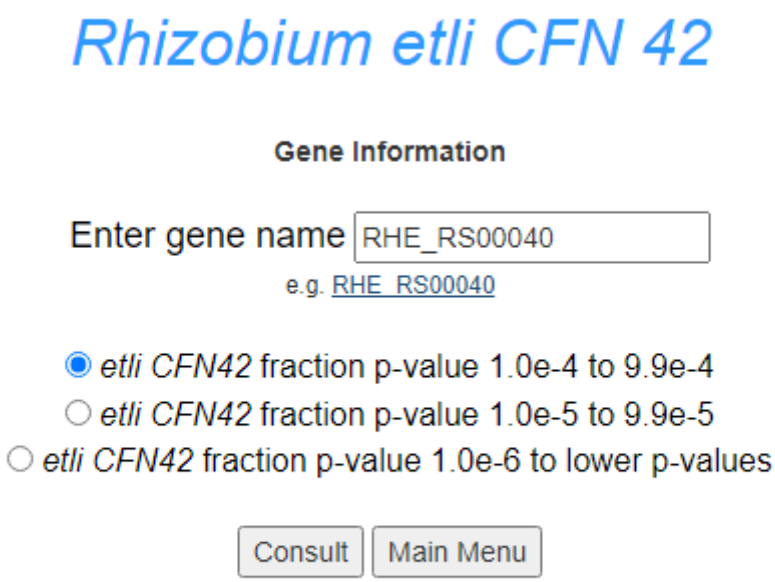
The motif search data is available by clicking on the "Download data from motifs" button, and data from graph by clicking on the "Download data from graph" button. The Cytoscape (Shannon P., *et al*, 2003) graph is available by clicking on the “View the network graph” button. Which can be rearranged by clicking on the circles representing the locus tags, it offers some layout formatting

options, and also the user can download the graph image by clicking on the "Save the image" button, then clicking on the graph with the right mouse button.
Network graph in Cytoscape with an “auto” option.



11.- Gene Information window: Provides a table with unique genes from motif information of the hypothetical regulon, there are some locus tag without information, below these, the numbers of nucleotide positions start and stop of gene sequence in the genome of genes sharing motifs of the query gene, strand location of the gene (+) or (-), Gene ID (GI), locus name, locus tag, protein product, length of the protein in aa, COG number, COG group, protein name, number for vicinity

and vicinity. The genes are neighbors if they are in one, two or three genes distance according to the gene numbering.



Enter a gene identifier.
Select a p-value, and click on the "Consult" button.
If there is not information of a gene query, a legend appears asking if the user wants to find information about function of genes with motifs in the query promoter gene. The result is provided in a table separated by tabs.

Gene information data.

entries

Start-stop	Chain	Gen ID	Locus	Locus tag	Protein product	Protein length (aa)	COG number	COG group	
RHE_RS08780									
RHE_RS31650									
RHE_RS00335									
RHE_RS11365									
295790..296740	-	24297161	nodI	RHE_RS30805	WP_016737487.1	316	COG1131	V	MULTIS
226819..227181	+	24297443	-	RHE_RS30445	WP_011053423.1	120			MULTIS
486351..487685	+	24302905	-	RHE_RS28690	WP_011428739.1	444	COG0161	H	aspartat
319840..320472	+	24303059	-	RHE_RS27920	WP_011428594.1	210	COG1309	K	TetR/Ac
180384..181676	+	24303179	-	RHE_RS27320	WP_011428479.1	430	COG1749	N	flagellar
175037..175876	-	24303183	-	RHE_RS27300	WP_011428475.1	279	COG0834	ET	ABC tra
14842..15837	-	24303344	-	RHE_RS26495	WP_042119940.1	331	COG0583	K	LysR fa
117367..118077	+	24302655	-	RHE_RS24540	WP_029531946.1	236	COG1802	K	MULTIS
115684..117162	-	24302656	-	RHE_RS24535	WP_011427958.1	492	COG1653	G	sugar A
2918..3457	-	24302754	-	RHE_RS24045	WP_011427863.1	179	COG1595	K	RNA po

Notice that there are not data available for some genes from the first column.

The genes grouped in COG are represented in the graph “Grouping of genes by COG”.

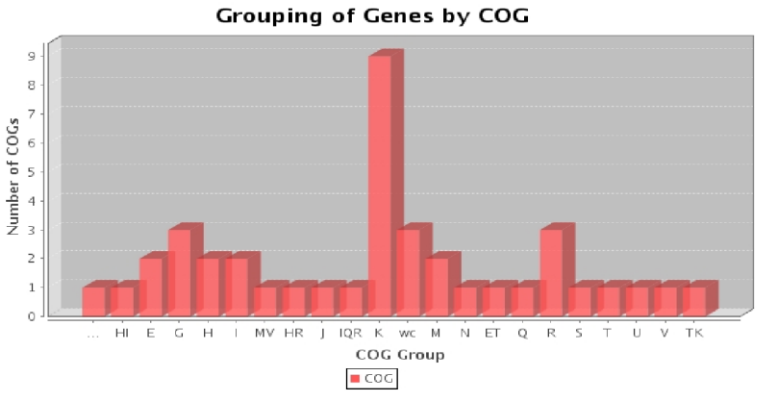
Rhizobium etli CFN 42

Gene Information
ID: RHE_RS00040
Pvalue: 1.0e-6 to lower values

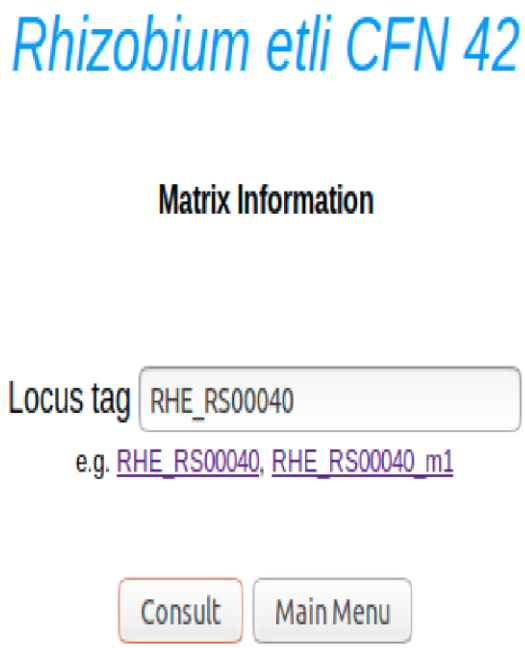
Return | Main Menu

Synonyms

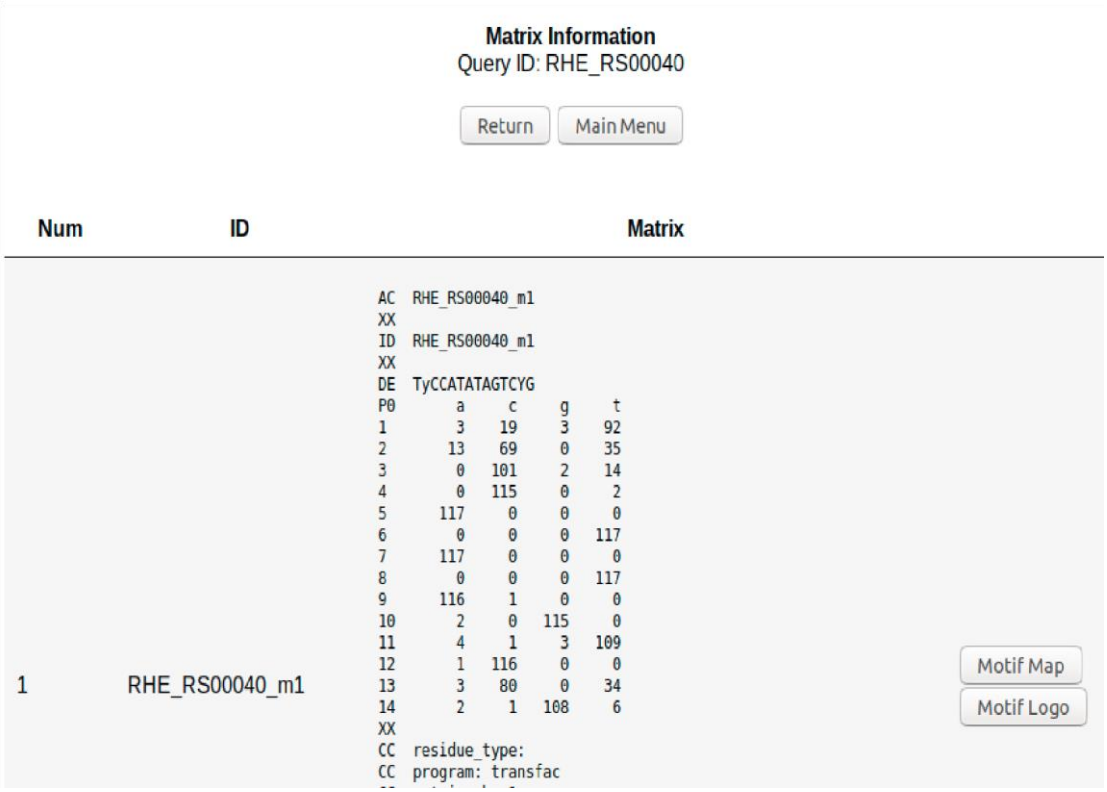
#1	#2	#3	#4
-	RHE_RS00040	WP_011423413.1	RHE_CH00007



12.- Matrices window. After selecting the strain in the main page, click on the Matrices button, all matrices of a gene are available by entering the locus tag or a particular matrix by typing the name of the matrix.



The matrix information appears after clicking on the “Consult” button, the matrix is in transfact format.



The Motif Map and the Motif Logo are available by clicking on the respective buttons.

13.- Example of how to use the matrix-scan data to search who potentially regulates a gene which is not a "TF-gene".

1.- Enter to the genome *Rhizobium etli* CFN42

2.- Enter to the Gene Information window

3.- Introduce the gene example (RHE_RS00040), select the p-value 1e-6 to lower p-values, press enter

4.- Download the corresponding file in a tsv format and open it with a gedit editor, make a copy and paste into an excel sheet.

You can see a table like this:

	Expected_regulon_RHE_RS00040											
Num	Start-stop	Chain	Gen_ID	Locus	Locus_tag	Protein_pro	Protein_leng	COG_number	COG_group	Protein_nam	Number_for	Vicinity
1	RHE_RS08780											
2	RHE_RS31650											
3	RHE_RS00335											
4	RHE_RS11365											
5	295790..2967	-	24297161	nodI	RHE_RS3080	WP_0167374	316	COG1131	V	MULTISPECIE	5922	
6	226819..2271	+	24297443	-	RHE_RS3044	WP_0110534	120			MULTISPECIE	5857	
7	486351..4876	+	24302905	-	RHE_RS2869	WP_0114287	444	COG0161	H	aspartate am	5522	
8	319840..3204	+	24303059	-	RHE_RS2792	WP_0114285	210	COG1309	K	TetR/AcrR fa	5372	

5.- Select and order the columns by the COG group, copy all the transcriptional regulators (COGK, nine transcriptional regulators).

	Expected_regulon_RHE_RS00040 ordered by de COG number column											
Num	Start-stop	Chain	Gen_ID	Locus	Locus_tag	Protein_pro	Protein_leng	COG_numbe	COG_group	Protein_nam	Number_for	Vicinity
17	143800..1453	+	24302144	-	RHE_RS2352	WP_0114277	525	COG0747	E	peptide ABC	4505	
27	3237413..323	+	24301416	-	RHE_RS1582	WP_0114263	367	COG0683	E	branched-ch	3042	
10	175037..1758	-	24303183	-	RHE_RS2730	WP_0114284	279	COG0834	ET	ABC transpo	5251	
13	115684..1171	-	24302656	-	RHE_RS2453	WP_0114279	492	COG1653	G	sugar ABC tra	4707	0

The Transcriptional regulators are:

RHE_RS27920
RHE_RS26495
RHE_RS24540
RHE_RS24045
RHE_RS17450
RHE_RS13800
RHE_RS08785
RHE_RS04880
RHE_RS04030

6.- Paste the transcriptional regulators in the left box of the application "Prediction of regulons" and type the RHE_RS00040 locus in the right box. Download the data. You see this output:

Locus_tag	Upstream_region	Matrix_ID	Chain	End_motif	Start_motif	Site	Weight	P_value	Ln_P_value	Significance
RHE_RS 00040	R	RHE_RS26495_m3	D	-73	-59	GATGAT TCAAATGAT	9.8	8.10E-06	-11.724	5.092

These data showed the RHE_RS26495_m3 matrix of the Transcriptional regulator RHE_RS26495 found a motif in the regulatory region of the RHE_RS00040 gene, which is not a TF-gene.

14.- Example of how to use the matrix-scan data to search how the transcriptional regulators may be interacting with the potential target genes.

1.- Select the genome *Rhizobium etli* CFN42

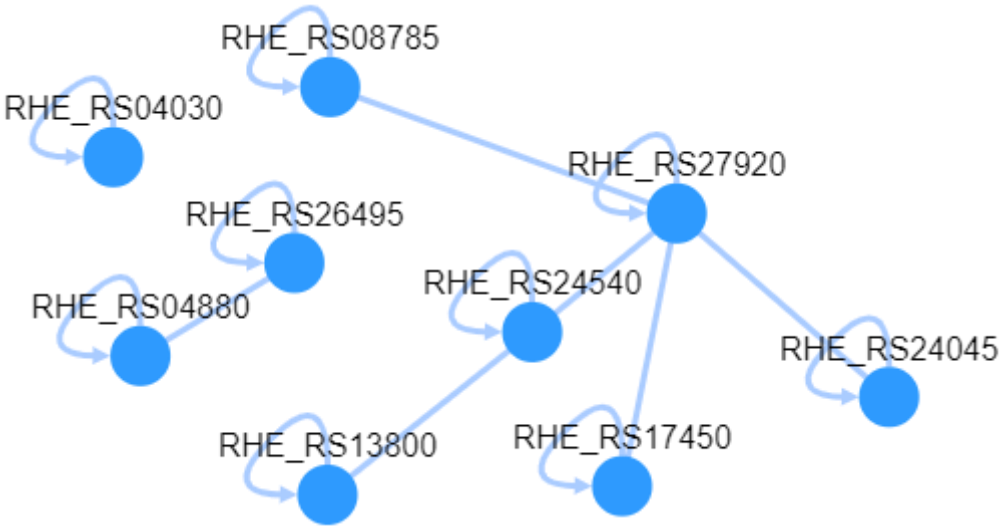
2.- Go to the Gene Information window

3.- Introduce the gene sample (RHE_RS00040), select the p-value 1e-6 to lower p-values

4.- Download the file in a tsv format and open it with a gedit editor, make a copy and paste in an Excel sheet as in the previous example

5.- Select and order the columns by the COG group, copy all the transcriptional regulators (nine transcriptional regulators) as in the previous example

6.- Paste the transcriptional regulators in both, the left and right boxes of the application “Prediction of regulons” click on see network graph.



This is a cytoscape network graph, it is an example of how the transcriptional regulators may be interacting from the expected regulon RHE_RS00040, with the option “auto”.

15.- Matrix-clustering. All the motifs represented in a matrix from only Transcription factors (TF´S) were grouped by its homology with the matrix-clustering program (Castro-Mondragon et al., Nucleic Acids Research (2017). This data show groups of TF´s sharing homology in their matrices,

consequently they potentially may be inter-regulated forming regulons, see **RhizoBindingSites v2.0 is a database of DNA motifs potentially involved in transcriptional regulation deduced from sites of the genome" Taboada-Castro at al., 2023.** Access is by clicking on the Matrix Clustering button, this will open a new window in their server. Once you see the "RSAT – matrix-clustering result" page, you have access to the data by clicking on the Logo Forest (dynamic browsing), Logo forest (rapid overview-low image quality), Clusters Summary, Individual Cluster View, Individual Motif View, Heatmap View and Additional Files (Castro-Mondragon et al., Nucleic Acids Research (2017), all the information is available. Since one TF may have more than one matrix, frecuently, a cluster is formed with matrices from the same gene. To see the cluster formed with more than one different gene, see the Supplementary Table 2.- Matrix-clustering_Analysis of the O_and_S_matrices of the above-mentioned manuscript.

16.- A Brief guide to analyze a genome in the RSAT web site http://embnet.ccg.unam.mx/rsat/.

In order to carry out an analysis of another genome, below is a brief introduction to the RSAT website http://embnet.ccg.unam.mx/rsat/ which is a specialized site to analyze a pre-charged genome.

For the analysis of a genome other that of Rhizobiales taxon, go to the RSAT web server located at http://embnet.ccg.unam.mx/rsat/, this site contains 56 programs for DNA sequences analysis (Nguyen et al., 2018), this site is made up of five distinct servers. For bacteria, you can choose RSAT-prokaryotes server, this site contains files for footprinting discovery algorithm like; the sequence of the bacterial genomes, the groups of orthologs, the upstream sequences and the background model. To get a footprint discovery algorithm, go to the page http://embnet.ccg.unam.mx/rsat/footprint-discovery_form.cgi. This page contains all the windows that the user needs to fill out and to submit the footprinting discovery algorithm, at the bottom of this program there is an example of how to complete this command by clicking on the "DEMO", it shows how to fill out the form to discover the motif in the promoter region of the E. coli *lexA* gene as an example, there is also a "sample Output" of the data, showing a file HTML output, click on the "Query" gen "*lexA*" link to obtain all the files with data used for the deduction of the *lexA* motif and the motif in logo format, as well as a footprint scan showing the conservation of the motifs discovered from *lexA* in members of the taxon Enterobacterales. This algorithm has the alternative of executing a single gene, a group of genes or a genome in the "Query genes" window. The deduced matrices are in the file with a ".tf" extension. The matrices are in transfac format, with these matrices you can run a matrix-scan analysis on the site http://embnet.ccg.unam.mx/rsat/matrix-scan-quick_form.cgi. With these simple steps you can analyze a pre-charged genome in the RSAT site just as we did for the RhizoBindingSites database.