



# RhizoBindingSites v2.0 Is a Bioinformatic Database of DNA Motifs Potentially Involved in Transcriptional Regulation Deduced From Their Genomic Sites

Bioinformatics and Biology Insights  
Volume 18: 1–10  
© The Author(s) 2024  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/11779322241272395



Hermenegildo Taboada-Castro<sup>1</sup>, Alfredo José Hernández-Álvarez<sup>1</sup> ,  
Jaime A Castro-Mondragón<sup>2</sup> and Sergio Encarnación-Guevara<sup>1</sup> 

<sup>1</sup>Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Mexico. <sup>2</sup>Centre for Molecular Medicine Norway, Nordic EMBL Partnership, University of Oslo, Oslo, Norway.

**ABSTRACT:** RhizoBindingSites is a *de novo* depurified database of conserved DNA motifs potentially involved in the transcriptional regulation of the *Rhizobium*, *Sinorhizobium*, *Bradyrhizobium*, *Azorhizobium*, and *Mesorhizobium* genera covering 9 representative symbiotic species, deduced from the upstream regulatory sequences of orthologous genes (O-matrices) from the Rhizobiales taxon. The sites collected with O-matrices per gene per genome from RhizoBindingSites were used to deduce matrices using the dyad-Regulatory Sequence Analysis Tool (RSAT) method, giving rise to novel S-matrices for the construction of the RhizoBindingSites v2.0 database. A comparison of the S-matrix logos showed a greater frequency and/or re-definition of specific-position nucleotides found in the O-matrices. Moreover, S-matrices were better at detecting genes in the genome, and there was a more significant number of transcription factors (TFs) in the vicinity than O-matrices, corresponding to a more significant genomic coverage for S-matrices. O-matrices of 3187 TFs and S-matrices of 2754 TFs from 9 species were deposited in RhizoBindingSites and RhizoBindingSites v2.0, respectively. The homology between the matrices of TFs from a genome showed inter-regulation between the clustered TFs. In addition, matrices of AraC, ArsR, GntR, and LysR ortholog TFs showed different motifs, suggesting distinct regulation. Benchmarking showed 72%, 68%, and 81% of common genes per regulon for O-matrices and approximately 14% less common genes with S-matrices of *Rhizobium etli* CFN42, *Rhizobium leguminosarum* bv. *viciae* 3841, and *Sinorhizobium meliloti* 1021. These data were deposited in RhizoBindingSites and the RhizoBindingSites v2.0 database (<http://rhizobindingsites.ccg.unam.mx/>).

**KEYWORDS:** Rhizobium, transcriptional regulation, motifs, matrices, regulon, binding sites

**RECEIVED:** April 16, 2024. **ACCEPTED:** July 12, 2024.

**TYPE:** Research Article

**FUNDING:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Part of this work was supported by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT-DUNAM), grant IN-213522 to S.E.-G.

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Sergio Encarnación-Guevara, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Av. Universidad s/n, Col. Chamilpa, Cuernavaca 62210, México. Email: [encarnac@ccg.unam.mx](mailto:encarnac@ccg.unam.mx)

## Introduction

The alphaproteobacteria and 2 species of beta-proteobacteria are able to establish symbiosis with leguminous plants.<sup>1</sup> When bacteria, transformed into a bacteroid form, are found within nodules in the roots of legume plants, they reduce atmospheric dinitrogen to ammonium via nitrogenase activity; this process is called symbiotic nitrogen fixation (SNF). The host plant provides the bacteroid with dicarboxylic acids to fuel the high demand for nitrogen fixation; correspondingly, the bacteroid provides this fixed nitrogen in the form of ammonium and amino acids to plant cells.<sup>2</sup> Intensified legume production based on biological nitrogen fixation instead of the predominant practice of using chemical fertilizers to cope with high food demand is a desirable and sustainable way to diminish the eutrophication of aquatic systems as well as emissions of nitrous oxide into the atmosphere.<sup>3–5</sup>

A better understanding of the role of bacteroid transcription factors (TFs) in SNF processes is fundamental for designing a better genetic regulatory circuitry to enhance the ability of symbiotic bacteria to reduce dinitrogen.<sup>6</sup> This requires great effort because *Rhizobium* is a free-living and symbiotic bacterium.

Regulatory circuits with intricate wiring have evolved to adapt to different environments,<sup>7,8</sup> and symbiotic types, which are carried out with different leguminous plants, showing various kinds of nodules (determined and undetermined).<sup>1</sup>

High-throughput approaches, such as transcriptomics, proteomics, and metabolomics, have been applied to study the symbiosis of symbiotic species.<sup>9</sup> Recently, metabolic maps of carbon-, nitrogen-, and phosphorus-integrating plant cells of legume nodules and bacteroids have been reported.<sup>1</sup> Moreover, a proteomic atlas of the host plant *Medicago truncatula* and its symbiont *Sinorhizobium meliloti* was constructed.<sup>6</sup> In addition, methods for annotating TF-binding sites and motif databases have been constructed, such as RegPrecise 3.0 (<https://regprecise.lbl.gov/>),<sup>10</sup> and RhizoBindingSites (see below) (<http://rhizobindingsites.ccg.unam.mx/>).<sup>11</sup> The next step would be to propose genetic circuits to build a transcriptional regulatory network for SNF. A bioinformatics method to construct *in silico* transcriptional regulatory networks, comparing the free-life and symbiosis in the maximal nitrogen fixation stage from *Rhizobium etli* CFN42 with the bean plant *Phaseolus vulgaris*, has been shown.<sup>8</sup> In addition, a computational reconstruction



of the transcriptional regulation of nitrogen fixation and signaling by oxygen in alphaproteobacteria was reported.<sup>12</sup> Although the main TFs have been identified experimentally, the lack of information on the functions of TFs related to carbon, phosphorus, and minerals, among others, makes it difficult to integrate them into a network. It is necessary to promote experimental designs that describe the role of a complete set of genes expressed in response to stimuli. The main goal of this work is to provide computational information on gene regulation for experimentalists to provide a more precise design of experiments.

We recently released the RhizoBindingSites database (<http://rhizobindingsites.ccg.unam.mx/>). These data were depurified (see the “Materials and Methods” section); the RhizoBindingSites database was obtained with the phylogenetic footprinting algorithm (Regulatory Sequence Analysis Tool [RSAT]) footprint discovery,<sup>13</sup> which aligns the promoters of orthologous genes for each genome in the order Rhizobiales to deduce a position-specific scoring matrix (PSSM), which represents the motif. These motifs are composed of spaced sites or dyads that are conserved among species and are potentially recognized by TFs. The dyads are 2 to 3 conserved nucleotide sequences spaced by a non-conserved sequence (separator) located in the regulatory region of the gen. The matrices, hereinafter referred to as “orthologue-derived-matrices” (O-matrices), were used to scan gene promoters in a genome to give rise to hypothetical regulons (h-regulons), which are defined as a group of genes sharing a motif of a TF, but, conventionally, non-TF genes also potentially have a motif in common. The output table (RhizoBindingSites, “Motif Information” section) contains an h-regulon per gene per genome with additional information.<sup>11</sup>

In this report, all the “sites” corresponding to the motifs of an h-regulon in the Motif Information section of RhizoBindingSites in the “sites” column (<http://rhizobindingsites.ccg.unam.mx/>) were used to newly predict matrices hereinafter referred to as “single-genome-matrices” (S-matrices). Remarkably, the RhizoBindingSites O-matrices originated from diverse orthologous genes from different species. In contrast, the S-matrices were deduced from the sites of the respective genomes. The S-matrices (as was shown for the O-matrices)<sup>11</sup> were used to scan the upstream –400 to –1 regulatory regions of all genes in their respective genomes, such as *R. etli* CFN 42, *R. etli* bv. *mimosae* Mim1, *Bradyrhizobium diazoefficiens* USDA 110, *Sinorhizobium fredii* NGR234, *S. meliloti* 1021, *R. l.* bv. *viciae* 3841, *Bradyrhizobium* sp. BTAi1, *Azorhizobium caulinodans* ORS 571, and *Mesorhizobium japonicum* MAFF303099, giving rise to RhizoBindingSites v2.0. We noticed that S-matrices contained more genes than O-matrices in their respective genomes. A comparison of logos showed that the frequency of nucleotide position specificity was better for S- than for O-matrices. However, for other motifs, the reduction of matrices yielded different nucleotide compositions.

A vicinity analysis of the genes per genome detected with S-matrices included more TFs than those detected with O-matrices. A matrix-clustering analysis of only matrices of TFs per genome for O- and S-matrices showed clusters of TFs with a minimum of 2 different genes, suggesting a functional relationship and different regulation for TFs of the same family.

## Materials and Methods

All bioinformatics methods used to construct the RhizoBindingSites database were performed using RSAT (<http://embnet.ccg.unam.mx/rsat/>) in a Linux environment.<sup>11</sup> A phylogenetic footprint discovery algorithm footprint discovery<sup>13</sup> was used, available on the web page RSAT, the guide for Users in the RhizoBindingSites database (<http://rhizobindingsites.ccg.unam.mx/>) and in (Appendix A).

### *Deduction of O-matrices in the RhizoBindingSites database*

PSSMs or O-motifs were deduced using the footprint discovery RSAT algorithm.<sup>14</sup> Briefly, this program receives, as input, the name of an organism, one or more gene names, and the name of a taxon (command in Appendix A). The program searched for orthologous genes in the given taxon for each gene of the desired organism with the best bidirectional hit and an E-value < 1.0e-5. For each selected ortholog, the program obtained the upstream sequences (–400 to –1) concerning the translation start site. These sequences were masked in redundant fragments of orthologs per gene. A redundant fragment is defined if it matches a previous segment of the same promoter set over at least 40 base pairs, with at most 3 substitutions. These purged sequences were used to detect overrepresented motifs with a motif discovery algorithm called dyad-analysis-RSAT.<sup>15,16</sup> This pattern-discovery program counts the number of occurrences of each dyad or tri nucleotides separated by 0 to 20 base pairs, also, short oligomers. The program dyad analysis assesses the significance of each dyad by comparing the observed occurrences in the orthologs of a gene with those expected by chance, according to a background model, taking all upstream sequences of all organisms belonging to the order Rhizobiales.<sup>11,16</sup> The “taxfreq” background model was used in this study,<sup>13</sup> where the prior probability of each dyad is estimated by computing the frequency observed for this dyad in the promoters of all genes of all organisms of the taxon. For each dyad, the risk of a false positive (nominal P-value) is computed using the binomial distribution as in Brohée et al.<sup>16</sup> The detected dyads or oligomers were assembled into PSSMs.<sup>17</sup>

### *Deduction of S-matrices in the RhizoBindingSites v2.0 database*

The sites located in the Motif Information section of the RhizoBindingSites database, created by the genomic matrix

scan with the filtered O-matrices from the RhizoBindingSites database, were extracted, and these nucleotide sequences representing the motifs were used as input to the program “create\_matrices\_from\_matches\_2018.pl” (Appendix A). This program generates 2 files: one with sites per predicted regulon and one with purged sites with the .fas extension in the new directory “Binding\_sites\_files.” This directory is used as input to the program “motifs\_discovery\_from\_matches\_sequences\_2018.pl” (Appendix B). This program deduces the matrices using a PSSM with the dyad-analysis method described previously,<sup>11</sup> creating a new directory called “motif\_discovery” that contains 1 sub-directory per h-regulon generated for each gene (the directory name is the same as the gene that gave rise to the regulon), which contain 3 files: dyads, the pattern assembly of the dyad, and 1 to 5 matrices in the transfac format with the .tf extension.

#### *Filtering of matrices*

The S-matrices were filtered by selecting those able to find motifs in their own promoter gene through the matrix-scan RSAT program at a  $P$ -value  $\leq 1e-4$ , as was shown for O-matrices.<sup>11,18,19</sup>

#### *Genomic matrix-scan for selecting targets with S-matrices*

The S-matrices were used to scan the  $-400$   $-1$  promoter region of all genes of their respective genome through the matrix-scan RSAT program. The upstream sequences were removed when overlapping, and the background model used was deduced with all upstream promoter regions from the taxon Rhizobiales with a  $P$ -value  $\leq 1e-4$  as was described.<sup>11</sup> These output data contained the h-regulons predicted with the S-matrices that were conventionally fractionated to stress the importance of having data with different levels of stringency: low ( $P$ -value:  $1.0e-4$  to  $9.9e-4$ ), medium ( $P$ -value:  $1.0e-5$  to  $9.9e-5$ ), and high ( $P$ -value:  $1.0e-6$  to lowest data value). Stringency is referred in the scanning process to the homology between the nucleotide sequence of the S-matrix and the sequence in the upstream regulatory region of genes in both the forward and reverse strands of DNA. These data define a hypothetical regulon (h-regulon) per gene, which is a group of genes in a genome detected during the scanning process with S-matrices of the aforementioned gene.<sup>11</sup> As matrices detect motifs in both DNA strands through matrix-scan analysis, additional *de novo* depuring step data were applied by selecting only targets in the codificant string of the target gene for O-(RhizoBindingSites) and S-matrices (RhizoBindingSites v2.0).

#### *Motif information and gene information sections*

Genomic matrix-scan analysis with the filtered S-matrices generated the motif information data similar to the MotifInformation

section of the RhizoBindingSites database. Gene Information sections were as in the RhizoBindingSites database.<sup>11</sup>

#### *Vicinity of genes from an h-regulon*

For bacterial genomes, proximal genes are often functionally related; this may occur because they may be in the same operon and regulated by the same TF,<sup>20-22</sup> but may also be genes with their promoters. Vicinity was defined as a group of genes at a distance of less than or equal to 3 genes in the genome.<sup>11</sup> The vicinity was searched for each gene in the genome and genes grouped in the COGK category<sup>23</sup> (Appendix C) related to transcriptional regulation, that is, TFs, response regulators, 2-component response regulators, sigma factors, and anti-sigma factors. At low, medium, and high levels of  $P$ -value stringency, as shown in the “Gene Information” section of the RhizoBindingSites database,<sup>11</sup> a comparison of the percentage of TFs with neighbors of h-regulons from the data of the O- and S-matrices at the 3  $P$ -values is shown (Supplementary Table 1C).

#### *Matrix clustering*

S- and O-matrices from only the COGK genes<sup>23</sup> per genome were selected for matrix-clustering RSAT analysis (Appendix D).<sup>24</sup> This program constructs clusters by grouping S- or O-matrices based on similarity. There is an output file “clusters\_motif\_names.tab” containing the clusters in a list (ie, cluster\_3 RHE\_RS06555\_m1, RHE\_RS06555\_m2, RHE\_RS03090\_m3, RHE\_RS03090\_m1, and RHE\_RS03090\_m2).

Genes RHE\_RS06555 and RHE\_RS03090 appeared repeatedly in this cluster. Unique genes per cluster were counted using the clusters\_motif\_names to avoid this redundancy.tab file and clusters with at least 2 different genes were extracted. Subsequently, NCBI information for each gene was added (Supplementary Table 2).

#### *Comparison of regulons from the Regprecise database, RhizoBindingSites, and RhizoBindingSites V2.0*

We referenced the Regprecise data because it included data from 56 regulons of the studied species in the RhizoBindingSites and RhizoBindingSites v2.0 databases. This was constructed using the experimental data from CoryRegNet 4.0, RegTransBase, and Regulon database (V6.0).<sup>10</sup> Briefly, data were obtained by searching for TFs and their target gene orthologs in a group of representative phylogenetically related species. Ortholog TFs with their target genes are called regulogs, and the search for regulogs was extended to the rest of the taxon species. Using their methodology, a position weight matrix was used to deduce the motifs for each regulator. The authors considered this data propagation to be accurate and conservative, indicating that it was not an attempt at automatic prediction.<sup>10</sup>

In contrast, our data were *ab initio* deduced from orthologous genes without a reference. First, the equivalent locus tags of the regulons of the extended section of Regprecise from *R. etli* CFN42, *Rhizobium leguminosarum* bv. *viciae* 3841, and *S. meliloti* 1021 were searched for compatibility with our locus tags (Supplementary Table 7A–C). Then, we used the application from our databases, “Prediction of regulatory networks,” by pasting the TF to the left box and the potential targets to the right box with the option “auto”; each of the 56 regulons was reported in the propagated section of Regprecise 3.0 (<https://regprecise.lbl.gov/>), for both O- and S-matrices (data not shown). As, in our data, the operon arrangement of the genes was not considered and most of the Regprecise regulons were operons, a paired list was constructed between the Regprecise operons and the genes found with the O- and S-matrices in our databases. Only the genes of regulons with genes of O- and S-matrices were considered common genes (Supplementary Table 7A–C).

### User's guide

RhizoBindingSites and RhizoBindingSites v2.0 are intuitive databases. In addition, a user's guide, a matrix-clustering window, and a synonyms converter application were included. For matrix-clustering data, given that a gene may have from 1 to 5 matrices, these matrices frequently cluster, giving rise to clusters from the same locus tag, which is low informative. To solve this, data with clusters containing more than 2 different genes may be consulted as a guide search (Supplementary Table\_2\_Matrix-clustering\_Analysis\_of\_O\_and\_S-Matrices). In addition, for the application “Prediction of regulatory Networks,” the user needs to use the locus tags proper of the application; to correct this, the synonyms converter was implemented (<http://rhizobindingsites.ccg.unam.mx/>). If the user needs to analyze a different bacterial species, there is an explanation in the user's guide.

## Results and discussion

### Statistics of RhizoBindingSites (O-matrices) and RhizoBindingSites v2.0 (S-matrices)

We showed the relevance of considering the stringency level of the inferred data on transcriptional regulation (see Materials and Methods). A comparative analysis of the average number of unique genes with O- and S-matrices for the 3 fractions of *P*-value showed 3871.11 and 3016.11 genes from the 9 genomes, respectively (Supplementary Table 1A). On average, there were 854.7 fewer genes with deduced S-matrices than O-matrices. Correspondingly, there were, on average, 76.27% and 65.68% of TFs with O- and S-matrices, respectively, concerning the total content of TFs in the 9 genomes (Supplementary Table 1A). These data indicated that there was 11% less TF content with S-matrices than with O-matrices

(Supplementary Table 1A), suggesting that sites from these O-matrices had low conservation, and it was challenging to find a consensus for the reduction of S-matrices. Unique genes detected per genome were determined after matrix-scan analysis of the respective genomes using these matrices. These data showed, on average, that, considering all data from low, medium, and high stringency, 5455.11 and 5542.55 unique genes were detected with the O- and S-matrices per 9 genomes, respectively. Furthermore, 81.63% and 82.91% of the genomic coverage concerned the average gene content of the 9 genomes. There was a 1.3% greater genomic coverage for the S-matrices than O-matrices (Supplementary Table 1B), although more genes with O-matrices were found. Given that the addition of genes due to the presence of operons was not considered in our data, the genomic coverage number could be more significant than these, showing essential genomic coverage with both O- and S-matrices.

### Matrices in a *transfac* format

As the next step is to uncover the genetic circuitry operating in a physiological condition, it is necessary to know the matrices of all the motifs of TFs known experimentally, that is, analysis is done to describe the matrices of the TFs for *Escherichia coli* K12 in the RegulonDB database.<sup>25</sup> Matrices are helpful in scanning all the upstream regulatory sequences of potential gene targets of the TFs in a genome. In addition, before constructing a network, it is advisable to conduct an analysis of homology between the matrices of a gen profile condition dependent on the matrix-clustering program.<sup>8</sup> To promote these, we are providing the O-matrices of 3187 TFs in RhizoBindingSites and S-matrices of 2754 TFs in RhizoBindingSites v2.0, which deserves a homology study with matrices from other ortholog TF genes from other studies (Supplementary Table 1A). As it is known, no-TF genes also have orthologs, and it is possible to deduce their matrices. Then, 34840 O-matrices that cover all the genes from the 9 species, including the matrices of the TFs and 27147 S-matrices from the same genomes, were deduced and deposited in RhizoBindingSites and RhizoBindingSites v2.0 databases, respectively. A multi-genomic matrix-clustering analysis of the TFs is, of course, necessary to know the conservation of the matrices between the orthologs from these species.

### S-matrices are more accurate than O-matrices

Transcription factors are frequently found behind gene-target neighbors.<sup>10,21,22</sup> We analyzed the vicinity of the h-regulons (see the Gene Information sections of RhizoBindingSites and RhizoBindingSites v2.0). Moreover, the percentage of TF content per *P*-value fraction per genome for neighboring genes was determined concerning the respective TF content per genome. At the *P*-value fractions of 1.0e-04, 1.0e-05, and

1.0e-06 for low, medium, and high stringent data, respectively, there were 7, 6, and 5 genomes with greater than 1.0% averages of TF content in the neighboring genes with S-matrices compared with O-matrices, respectively. Genomes *R. etli* CFN42, *R. etli* bv. *mimosae* str. Mim1, *S. meliloti* 1021, *Bradyrhizobium* sp. BTAi1 and *A. caulinodans* ORS571 showed, for S-matrices, a greater TF content in the neighboring genes in the 3 *P*-value ranges. *Rhizobium leguminosarum* bv. *viciae* 3841 showed a more significant TF number with neighboring genes in the 1.0e-04- and 1.0e-06-*P*-value ranges. In contrast, *S. fredii* NGR234 showed a greater TF content with neighboring genes in the 1.0e-04- and 1.0e-05-*P*-value ranges (Supplementary Table 1C).

Although, on average, 11% fewer TFs with S-matrices than with O-matrices (see above) (Supplementary Table 1A), most of the genomes showed a greater TF content of the neighboring genes detected with S-matrices than with O-matrices (Supplementary Table 1C). The homology between the nucleotide sequences of the S-matrices and the upstream regulatory regions in the respective genomes was higher than that with the O-matrices, corresponding to the deduction of S-matrices with their own genomic sites. These data suggest that the S-matrices show greater accuracy than the O-matrices.

#### Clusters with O- and S-matrices

The homology of the O- and S-matrices of TFs per genome was analyzed separately using a matrix-clustering program.<sup>14,24</sup> Homology between the TF-matrices was expected because of the functional relationship observed in an *E. coli* K-12 transcriptional regulatory network.<sup>26</sup> In addition, the hierarchy of TFs of minimal medium growth and symbiotic proteomes from *R. etli* CFN42 has been shown.<sup>8</sup> These homologies were displayed in hierarchical dendrograms using the HCLUST algorithm.<sup>24</sup> Output matrix-clustering data with O- and S-matrices are available at RhizoBindingSites and RhizoBindingSites v2.0, respectively (<http://rhizobindingsites.ccg.unam.mx/>). On the main page, there is a section called “Matrix-clustering,” which opens an archive.html that displays a new web page with some sections (dendrograms of the corresponding O- and S-matrices per genome) which are available by clicking on the “Logo Forest (dynamic browsing)” section, providing 18 directories (<http://rhizobindingsites.ccg.unam.mx/>). TFs with clustered O- and S-matrices were, on average, 61.8% and 59.3%, respectively, and the average of clusters were 206 and 184, respectively (Supplementary Table 1D–E). According to the average of TFs with O- and S-matrices (Supplementary Table 1A), clustered matrices for O- and S-matrices represented 14.47% and 6.38% fewer TFs, respectively (Supplementary Table 1D–E). These data show that there is more significant homology between the S-matrices than between the O-matrices. Data from 18 matrix-clustering analyses were extracted to avoid the redundancy of genes per cluster; clusters

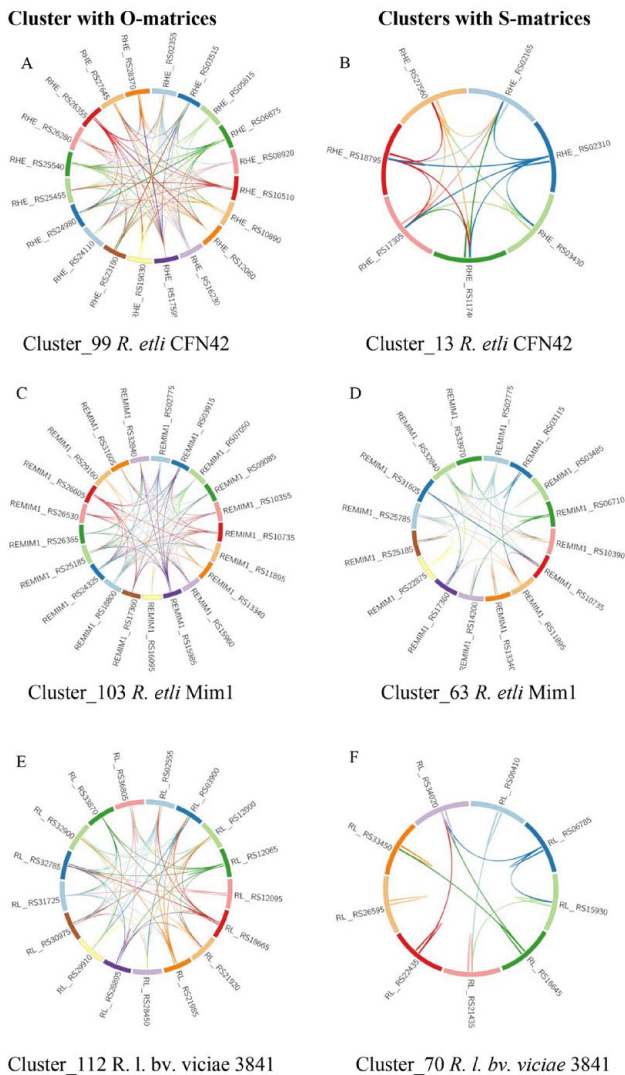
between other genes also contained genes with the same functions (Supplementary Table 2). These genomes contained more than 10% of genes with functional redundancy, that is, in *R. etli* CFN42, there were 15 identified AraC genes with deduced matrices (Supplementary Table 2) (see below Supplementary Table 3). Because an interrelationship between TFs is expected, it is essential to know the functionality of O- and S-matrices by their ability to form regulons from clustered TFs.

#### Clustered TFs may be functionally related

We have previously shown that genes clustered with a TF or TFs are potentially functionally organized in a hypothetical regulon (h-regulon).<sup>8</sup> Some clusters were analyzed in the application from RhizoBindingSites and RhizoBindingSites v2.0 (“Prediction of Regulatory Networks”) by pasting the TF genes of a cluster from Supplementary Table 2 to both boxes; as regulators and as targets, a medium restriction level of 1.0e-05 was selected, and TF-target data were used to design graphs in the Circos program.<sup>27</sup> These data showed the TF genes of clusters with O- and S-TF-matrices are related; that is, cluster\_99 and cluster\_13 were for the O- and S-TF-matrices from *R. etli* CFN42, respectively. Cluster\_103 and cluster\_63 from *R. etli* Mim1 were for the O- and S-TF-matrices, respectively. Cluster\_112 and cluster\_70 from *R. leguminosarum* bv. *viciae* 3841 are for the O- and S-TF-matrices (Figure 1). Cluster\_99 and cluster\_19 were for O- and S-TF-matrices from *S. meliloti* 1021, respectively (Supplementary Continued Figure 1). If there is no relationship between TF-TF genes, the gene appears isolated, showing only that it recognizes a motif in its promoter, that is, for cluster\_63 from *R. etli* bv. *Mimosae* Mim1, gene RHEMIM1\_RS034085, cluster\_70 from *R. leguminosarum* bv. *viciae* 4841, and gene RL\_RS26595 (Figure 1). Bioinformatics methods include false-positive data, and a frontier challenge of bioinformatics sciences is to construct methods to diminish these data. A method was proposed to construct transcriptional regulatory networks, lowering low-stringency data with a matrix-clustering method, favoring TF gene-target relationships with conserved motifs in both the TF and gene target of the same cluster.<sup>8</sup> Therefore, the quality of the matrices is crucial for constructing regulons. Matrix clustering of TFs provides information for constructing a global transcriptional regulatory network per genome.

#### Nucleotide composition of O- and S-matrix logos

The re-definition of matrices from sites of the genome is a novel strategy not registered in the literature. A total of 33 logos of the same gene were selected to compare the nucleotide compositions of the O- and S-matrices (Figure 2). It should be noted that cluster\_83 O-matrix RHE\_RS14875\_m1 from *R. etli* CFN42 has a nucleotide sequence AAATTG, whereas in cluster\_106, the S-matrix RHE\_RS14875\_m5 was replaced



**Figure 1.** Clusters of matrices of TF genes formed regulons in the application “Prediction of regulatory networks” for O-matrices in RhizoBindingSites and S-matrices in RhizoBindingSites v2.0. Circos graphs from clusters A, C, and E and in Supplemental Continued Figure 1: G, I, K, M, O, and Q are with O-matrices. Clusters B, D, and F and in Supplemental Continued Figure 1: H, J, L, N, P, and R, are with S-matrices. Representative clusters of TFs from each species were selected for having a high number of unique genes.

with AAATAT, and the sequence ACAATTT was present in both O- and S-matrices. Similarly, in cluster\_51, the O-matrix RL\_RS29140\_m5 from *R. leguminosarum* bv. *viciae* 3841 had the sequence AAAGTGTATGCAA, as in cluster\_288 S-matrix RL\_RS29140\_m1, but with a greater frequency in the S-matrix. In contrast, the GGACGTGCCA sequence in RL\_RS29140\_m5 was replaced with TTTTCG in RL\_RS29140\_m1 (Figure 2).

Interestingly, in cluster\_261, O-matrix REMIMI\_RS13960\_m4 from *R. etli* bv. *mimosae* Mim1, the sequence GATC-30-GATC was present, whereas, in cluster\_41, the S-matrix REMIMI\_RS13960\_m1, the oligo TTGCAG GATCGTGCAA was found, which included the GATC sequence; however, the spacing and double GATC sequence

was not conserved in the S-matrix site (Figure 2). For cluster\_286, the O-matrix RHE\_RS17145\_m5 from *R. etli* CFN42, and cluster\_69, the S-matrix RHE\_RS17145\_m5 showed the sequence ACGAATAAT, but with a greater frequency in the S-matrix RHE\_RS17145\_m5. Moreover, the O-matrix showed the oligo GGCATCACT, which is present in the orthologs of RHE\_RS17145 in the *Rhizobiales* taxon. Consequently, this oligo was present in the O-matrix but was not a consensus in the sites of the *R. etli* CFN42 genome (Figure 2).

From these data, we observed that some nucleotides present in the O-matrices were absent in the S-matrices. There is a re-definition of the nucleotides of the logos in the S-matrices, and the S-matrices are more suitable for their respective genomes. Around 23 cases with S-sites exhibited conserved nucleotide composition with O-matrices but with a greater frequency than those in the O-matrices (Figure 2 and Supplementary Continued Figure 2). In addition, there was a drastic re-composition of nucleotides in the logos, that is, cluster\_286 and cluster\_69 (Figure 2), cluster\_464 O-matrix *M. japonicum* MAFF303099 MAFF\_RS10245\_m2, and cluster\_146 S-matrix *M. japonicum* MAFF303099 MAFFRS10245\_m2 (Supplementary Continued Figure 2). In addition, cluster\_179 included the O-matrix *R. etli* CFN42, RHE\_RS27560\_m5, and cluster\_13 included the S-matrix *R. etli* CFN42, RHE\_RS27560\_m4, cluster\_180 O-matrix *R. leguminosarum* bv. *viciae* 3841 RL\_RS19380\_m5 and cluster\_135 S-matrix *R. l. bv. viciae* 3841 RL\_RS19380\_m3 (Supplementary Continued Figure 2). Moreover, cluster\_721 O-matrix *R. etli* CFN42 RHE\_RS23180\_m3 and cluster\_65 S-matrix *R. etli* CFN42, RHE\_RS23180\_m5 (Supplementary Continued Figure 2). In addition, cluster\_14 O-matrix *S. fredii* NGR234 NGR\_c12400\_m5 and cluster\_52 S-matrix *S. fredii* NGR234 NGRc12400\_m2 (Supplementary Continued Figure 2). This indicates the sites obtained with the O-matrices had a lax consensus. Consequently, the rededuction of new S-matrices from these sites was with a low consensus site, generating these differences. We notice when AraC TFs from *R. etli* CFN42 cluster\_721 O-matrix RHE\_RS23180\_m3 and cluster\_318 O-matrix RHE\_RS15070\_m1 were compared (Figure 2). In addition, their corresponding cluster\_65 S-matrix RHE\_RS23180\_m5 and cluster\_205 S-matrix RHE\_RS15070\_m2 motifs (Supplementary Continued Figure 2) showed different logos despite belonging to the same family. This data suggested that 2 AraC TFs are differentially regulated. To determine the generalizability of this observation, we analyzed more families of TFs.

#### *Potentially different regulation of the TFs of the AraC, ArsR, GntR, and LysR families*

*AraC* family. The 2 AraC TFs were clustered in different groups because of the different nucleotide compositions of their motifs, suggesting distinct transcriptional regulation (see above). Matrix clustering of the O- and S-matrices of AraC



**Figure 2.** Nucleotide composition of motifs from O- and S-matrices of the same TF was compared. The frequency of some nucleotides position-specific was greater for motifs of S- than O-matrices. For other cases, re-definition of nucleotides from motifs of S-matrices as compared with O-matrices was observed, while, for other motifs, a completely new composition of nucleotides was observed.

TFs from *R. etli* CFN42 and *R. leguminosarum* bv. *viciae* 3841 and *S. meliloti* 1021 species were identified (Supplementary Table 3A–C). Only the AraC genes with deduced O- and S-matrices were considered. Furthermore, only clusters containing more than 2 different genes were considered. Note that

a gene may appear more than once in the same cluster because each gene may have 1 to 5 matrices; if these matrices are closely homologous, they are grouped in the same cluster (RhizoBindingSites, “Matrix Clustering” section).<sup>24</sup> *Rhizobium etli* CFN42 contained 15 unique AraC TFs with deduced matrices. Only

cluster\_99 contained the genes RHE\_RS1795 and RHE\_RS23180; the other 13 AraC TFs were in different clusters. This indicated that 14 different matrices covered the 15 AraC TFs in this dataset. In contrast, in the S-matrices, the same AraC TFs in cluster\_99 were grouped in cluster\_65. An additional cluster\_12 with the RHE\_RS11860, RHE\_RS32555, and RHE\_RS11860 AraC TFs was shown (Supplementary Table 3A), and the other 12 AraC TFs were grouped into different clusters, suggesting that 14 distinct matrices are involved in the transcriptional regulation of these genes (Supplementary Table 3A). *Rhizobium leguminosarum* bv. *viciae* 3841 with O- and S-matrices had 15 unique AraC TFs. For the O-matrices, cluster\_112 contained RL\_RS28895 and RL\_RS30975 AraCs, whereas cluster\_125 contained the RL\_RS06995 and RL\_RS16545 AraC TFs, which were equally clustered with the S-matrices in cluster\_14 and cluster\_46, respectively. A total of 11 other AraC TFs were grouped into different clusters, suggesting that 13 different motifs regulate these genes (Supplementary Table 3B). Moreover, for *S. meliloti* 1021, 4 unique AraC TFs with both O- and S-matrices were found. All the AraC TFs were grouped into clusters for the O-matrices. For S-matrices, cluster\_3, containing the SMA1454 and SMA2163 AraC TFs, were identified (Supplementary Table 3C). Duplication of genes is frequent in species of the Rhizobiales taxon, that is, in *R. etli* CFN42, operons *nifHDK*, *FixNOQP*, and *FixGHIS*.<sup>28,29</sup> TFs of the same family may be involved in different metabolic tasks,<sup>30</sup> which would explain why they are potentially expressed in different metabolic conditions, thus with a different transcriptional regulation, and only some genes with highly homologous matrices may co-occur in their expression.<sup>8</sup> An identical analysis of the matrix-clustering data of the ArsR, GntR, and LysR families with O-matrices was performed to determine whether members of TFs of the same family were clustered in different groups.

**ArsR family.** *R. etli* CFN42, and *R. leguminosarum* bv. *viciae* 3841 and *S. meliloti* 1021 contained 8, 5, and 5 unique ArsR TFs, respectively (Supplementary Table 4A–C). *Rhizobium etli* CFN42 belonged to cluster\_172 with RHE\_RS05330 and RHE\_RS05495 ArsR TFs, and the last 6 ArsR TFs belonged to different clusters. In contrast, the ArsR TFs were not grouped in the same cluster as *R. leguminosarum* bv. *viciae* 3841 and *S. meliloti* 1021, meaning that they have different motif sequences (Supplementary Table 4A–C).

**GntR family.** For the GntR family from *R. etli* CFN42, 5 clusters were shown (Supplementary Table 5A): cluster\_123 with RHE\_RS10960 and RHE\_RS22955; cluster\_3 with RHE\_RS23065 and RHE\_RS28665; cluster\_344 with RHE\_RS10960 and RHE\_RS24540; cluster\_60 with RHE\_RS24620 and RHE\_RS27280; and cluster\_83 with RHE\_RS14875, RHE\_RS24620, RHE\_RS04625, RHE\_RS10960, and RHE\_RS29975. A total of 10 different GntR TFs were then

grouped into these 5 clusters, and the last 4 GntR TFs, from a total of 14, were each grouped into different clusters (Supplementary Table 5A). For *R. leguminosarum* bv. *viciae* 3841, cluster\_108 had 3 genes, and RL\_RS35915, RL\_RS15070, and RL\_RS26800, and cluster\_137 with RL\_RS15070 and RL\_RS29140 were found, totaling 4 different TFs; then, 8 TFs from 12 unique TFs were found in different clusters (Supplementary Table 5B). In *S. meliloti* 1021, cluster\_26 contained the SMA0160 and SMA0062 GntR TFs, indicating that 7 of the 9 unique TFs were located in different clusters (Supplementary Table 5C).

**LysR family.** The most abundant family of TFs was the LysR family in *E. coli* k-12.<sup>31,32</sup> Matrix-clustering analysis with O-matrices for LysR TFs from *R. etli* CFN42 revealed 15 clusters grouping 27 genes from 42 unique LysR TFs with matrices. The remaining 15 LysR TFs were located in different clusters (Supplementary Table 6A). In *R. leguminosarum* bv. *viciae* 3841, 24 genes were distributed in 15 clusters, and the last 22 of 46 LysR TFs with matrices were located in different clusters (Supplementary Table 6B). Furthermore, for *S. meliloti* 1021, 7 clusters contained 13 LysR TFs, and 11 from a total of 24 LysR TFs were grouped into different clusters (Supplementary Table 6C). These data showed that, for O-matrices, TFs from the same family were grouped into clusters potentially subjected to different transcriptional regulations. Consequently, TFs with the same transcriptional regulation should be grouped within the same cluster. For S-matrices in general, a lower number of groups of clusters was found than with O-matrices, meaning that S-matrices were more different than the O-matrices.

#### Comparison of O- and S-matrix data with the Regprecise data

To determine the confidence of our data, a comparison of 57 regulons reported in the section on propagated data from Regprecise<sup>10</sup> and the corresponding h-regulons obtained with the O-matrices in RhizoBindingSites and S-matrices in RhizoBindingSites v2.0 databases were done. In the data on the 57 regulons, the authors did not determine whether the regulons were operons; the numeration of the locus tags and the assignment of the locus names strongly suggest that they were operons. Considering *a priori*, that they are operons, *R. etli* CFN42 (Supplementary Table 7A), *R. leguminosarum* bv. *viciae* 3841 (Supplementary Table 7B), and *S. meliloti* 1021 (Supplementary Table 7C) showed 72.14%, 68.07%, and 81.34% of common genes in the regulons with O-matrices (see the summary in Supplementary Table 7D); meanwhile, 58.96%, 52.10%, and 65.62% of common genes were with S-matrices, respectively. These data show that our predictions coincide with the data from the Regprecise database. Note that these data only have O- and S-matrices that recognize a motif in the



upstream regulatory region of TFs and are expected to be autoregulated TFs. Instead, we noticed that 22.8% of the TFs of regulons from Regprecise were not potentially autoregulated (YrdX, NadQ, NrdR, HutC, Mur, ModE, NifA, HrcA, RhiR, CadR-PbrR, SMc04260, AnsR, and SMb20039); this likely limits the detection of all other genes of regulons from Regprecise. Also, these data showed 13.17%, 15.97%, and 15.72% fewer common genes detected in the S-matrices than the O-matrices for *R. etli* CFN42 (Supplementary Table 7A), *R. leguminosarum* bv. *viciae* 3841 (Supplementary Table 7B), and *S. meliloti* 1021 (Supplementary Table 7C), respectively (see summary in Supplementary Table 7D). Therefore, an additional analysis to determine the number of genes that a TF potentially regulates (TFs hierarchy) in O- and S-matrices for *R. etli* CFN42, *R. leguminosarum* bv. *viciae* 3841, and *S. meliloti* 1021 showed that, at a *P*-value of 1.0e-04, 359, 545, and 445, fewer genes were detected on average with S- than O-matrices, respectively. At a *P*-value of 1.0e-05, on average, 58, 93, and 82 fewer genes were detected in the S- than with the O-matrices, respectively. Meanwhile, at a *P*-value of 1.0e-06, on average, 6, 9, and 10 fewer genes were detected with S- than with O-matrices, respectively (data not shown). These data agree with the lower detection rate of genes with S-matrices than those with O-matrices of regulons from Regprecise. The difference from the data showing a greater number of unique genes detected with S- than O-matrices (see above) (Supplementary Table 1B) is that, for this analysis, unique genes were considered separately at *P*-values of 1.0e-04, 1.0e-05, and 1.0e-06, instead of all of them being considered together.

Altogether, these data showed that the consensus of the matrices is a determinant for the accuracy of the predictions; as was aforementioned, a lax consensus promoter of genes<sup>33</sup> determined a low consensus of O-matrices and consequently, the S-matrices, also the low number of orthologous genes affects the quality of the deduced O-matrices. Recently, a bioinformatic study to infer regulons by searching orthologs from both the TF and their target genes from experimentally determined data was done, meaning that there is a core of targets per TF,<sup>34</sup> this is highly conservative data, and there are no motifs for this inferred regulons, making this data unappropriated to analyze accuracy.

## Conclusions

In the face of global warming, increasing biological constraints are imposed on food production, and the engineering of metabolic pathways to achieve more efficient biological nitrogen fixation in the symbiosis between the Rhizobiales taxon species and their respective host leguminous plants is a desirable strategy. With the availability of genomic sequences, bioinformatics methodologies are essential for extracting pertinent information on transcriptional regulation at the genomic level. Sites,

which are conserved short nucleotide sequences located in the upstream regulatory region of genes called motifs, potentially involved in transcriptional regulation, were obtained with the O-matrices deposited in the RhizoBindingSites database. These sites were used to re-deduce new S-matrices. Pointing O-matrices were deduced from the upstream sequences of the orthologous genes of each gene per genome, and S-matrices were deduced from the sites of the genome obtained with the O-matrices. Although fewer TF genes had S-matrices than O-matrices, a genomic scan analysis with both O- and S-matrices showed that S-matrices had a 1% greater genomic coverage than O-matrices. Globally, these data demonstrate that sequences of S-matrices have more homology with the upstream regulatory sequences of genes than O-matrices in the corresponding genome. Genes in the vicinity detected with S-matrices had a greater TF content than those detected with O-matrices. A hierarchical functional interrelationship was inferred between the TFs.<sup>8</sup> Hypothetical regulons were formed with TFs grouped using a matrix-clustering method. In addition to this functional validation of the O- and S-matrices, the deduced regulons represented the simplest structure of a transcriptional regulatory network, thus opening the window for the conception of a global transcriptional regulatory network.

This knowledge of the conservation of motifs from symbiotic species, potentially involved in transcriptional regulation, will allow for better experiment designs to decipher how wiring occurs in a network.

## Acknowledgements

We wish to thank the administration of technology information UATI group from the CCG UNAM for the support in the Information technology and equipment. Also, we also like to thank the Bioinformatic analysis UAB group from the CCG UNAM for their support with bioinformatic methods.

## Author Contributions

HT-C: conceptualization, methodology, software, validation, investigation, data curation, writing original draft, and visualization. AJH-A: software, resources, data curation, and visualization. JAC-M: conceptualization, methodology, and software. SE-G: conceptualization, validation, resources, writing review & editing, supervision, project administration, and funding acquisition.

## ORCID iDs

Alfredo José Hernández-Álvarez  <https://orcid.org/0009-0006-1025-9931>

Sergio Encarnación-Guevara  <https://orcid.org/0000-0001-7889-1681>

## SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

## REFERENCES

- Liu A, Contador CA, Fan K, Lam HM. Interaction and regulation of carbon, nitrogen, and phosphorus metabolisms in root nodules of legumes. *Front Plant Sci.* 2018;9:1860. doi:10.3389/fpls.2018.01860
- Udvardi MK, Day DA. Metabolite transport across symbiotic membranes of legume nodules. *Annu Rev Plant Physiol Plant Mol Biol.* 1997;48:493-523. doi:10.1146/annurev.arplant.48.1.493
- Blesh J. Feedbacks between nitrogen fixation and soil organic matter increase ecosystem functions in diversified agroecosystems. *Ecol Appl.* 2019;29:e01986. doi:10.1002/eap.1986
- Erisman JW, Galloway J, Seitzinger S, Bleeker A, Butterbach-Bahl K. Reactive nitrogen in the environment and its effect on climate change. *Curr Opin Environ Sustain.* 2011;3:281-290. doi:10.1016/j.cosust.2011.08.012
- Ferguson BJ, Mens C, Hastwell AH, et al. *Legume Nodulation: The Host Controls the Party*. Vol. 42. Blackwell Publishing; 2019:41-51. doi:10.1111/pcc.13348
- Marx H, Minogue CE, Jayaraman D, et al. A proteomic atlas of the legume *Medicago truncatula* and its nitrogen-fixing endosymbiont *Sinorhizobium meliloti*. *Nat Biotechnol.* 2016;34:1198-1205. doi:10.1038/nbt.3681
- Ibarra-Arellano MA, Campos-González AI, Treviño-Quintanilla LG, Tauch A, Freyre-González JA. Abasy Atlas: a comprehensive inventory of systems, global network properties and systems-level elements across bacteria. *Database.* 2016;2016:baw089. doi:10.1093/database/baw089
- Taboada-Castro H, Gil J, Gómez-Caudillo L, Escorcía-Rodríguez JM, Freyre-González JA, Encarnación-Guevara S. *Rhizobium etli* CFN42 proteomes showed isoenzymes in free-living and symbiosis with a different transcriptional regulation inferred from a transcriptional regulatory network. *Front Microbiol.* 2022;13:947678. doi:10.3389/fmicb.2022.947678
- Lardi M, Pessi G. Functional genomics approaches to studying symbioses between legumes and nitrogen-fixing rhizobia. *High Throughput.* 2018;7:15. doi:10.3390/ht7020015
- Novichkov PS, Kazakov AE, Ravcheev DA, et al. RegPrecise 3.0: a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics.* 2013;14:745. doi:10.1186/1471-2164-14-745
- Taboada-Castro H, Castro-Mondragón JA, Aguilar-Vera A, Hernández-Álvarez AJ, van Helden J, Encarnación-Guevara S. RhizoBindingSites, a database of DNA-binding motifs in nitrogen-fixing bacteria inferred using a footprint discovery approach. *Front Microbiol.* 2020;11:567471. doi:10.3389/fmicb.2020.567471
- Tsoy OV, Ravcheev DA, Čuklína J, Gelfand MS. Nitrogen fixation and molecular oxygen: comparative genomic reconstruction of transcription regulation in Alphaproteobacteria. *Front Microbiol.* 2016;7:1343. doi:10.3389/fmicb.2016.01343
- Janky R, van Helden J. Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution. *BMC Bioinformatics.* 2008;9:37. doi:10.1186/1471-2105-9-37
- Nguyen NTT, Contreras-Moreira B, Castro-Mondragón JA, et al. RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Res.* 2018;46:W209-W214. doi:10.1093/nar/gky317
- van Helden J, Rios AF, Collado-Vides J. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.* 2000;28:1808-1818. Accessed April 20, 2017. <http://www.ncbi.nlm.nih.gov/pubmed/10734201>
- Brohée S, Janky R, Abdel-Sater F, Vanderstocken G, André B, van Helden J. Unraveling networks of co-regulated genes on the sole basis of genome sequences. *Nucleic Acids Res.* 2011;39:6340-6358. doi:10.1093/nar/gkr264
- Santana-García W, Castro-Mondragón JA, Padilla-Gálvez M, et al. RSAT 2022: regulatory sequence analysis tools. *Nucleic Acids Res.* 2022;50:W670-W676. doi:10.1093/NAR/GKAC312
- Defrance M, Janky R, Sand O, van Helden J. Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nat Protoc.* 2008;3:1589-1603. doi:10.1038/nprot.2008.98
- Thomas-Chollier M, Sand O, Turatsinze JV, et al. RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.* 2008;36:W119-W127. doi:10.1093/nar/gkn304
- Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.* 2004;5:R35. doi:10.1186/gb-2004-5-5-r35
- Pannier L, Merino E, Marchal K, Collado-Vides J. Effect of genomic distance on coexpression of coregulated genes in *E. coli*. *PLoS ONE.* 2017;12:e0174887. doi:10.1371/journal.pone.0174887
- Strong M, Mallick P, Pellegrini M, Thompson MJ, Eisenberg D. Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach. *Genome Biol.* 2003;4:R59. doi:10.1186/gb-2003-4-9-r59
- Tatusov RL, Fedorova ND, Jackson JD, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics.* 2003;4:41. doi:10.1186/1471-2105-4-41
- Castro-Mondragón JA, Jaeger S, Thieffry D, Thomas-Chollier M, van Helden J. RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res.* 2017;45:e119-e119. doi:10.1093/nar/gkx314
- Tierrafria VH, Rioualen C, Salgado H, et al. RegulonDB 11.0: comprehensive high-throughput datasets on transcriptional regulation in *Escherichia coli* K-12. *Microb Genom.* 2022;8:mgen000833. doi:10.1099/MGEN.0.000833
- Martínez-Antonio A, Collado-Vides J. Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol.* 2003;6:482-489. doi:10.1016/J.MIB.2003.09.002
- Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* 2009;19:1639-1645. doi:10.1101/gr.092759.109
- Valderrama B, Dávalos A, Girard L, Morett E, Mora J. Regulatory proteins and cis-acting elements involved in the transcriptional control of *Rhizobium etli* reiterated *nifH* genes. *J Bacteriol.* 1996;178:3119-3126.
- Granados-Baeza MJ, Gómez-Hernández N, Mora Y, Delgado MJ, Romero D, Girard L. Novel reiterated Fnr-type proteins control the production of the symbiotic terminal oxidase *cbb3* in *Rhizobium etli* CFN42. *Mol Plant Microbe Interact.* 2007;20:1241-1249. doi:10.1094/MPMI-20-10-1241
- Cortés-Avalos D, Martínez-Pérez N, Ortiz-Moncada MA, et al. An update of the unceasingly growing and diverse AraC/XylS family of transcriptional activators. *FEMS Microbiol Rev.* 2021;45:1-13. doi:10.1093/femsre/fuab020
- Pérez-Rueda E, Collado-Vides J. The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res.* 2000;28:1838-1847. Accessed May 2, 2017. <http://www.ncbi.nlm.nih.gov/pubmed/10734204>
- Pérez-Rueda E, Hernández-Guerrero R, Martínez-Núñez MA, Armenta-Medina D, Sánchez I, Ibarra JA. Abundance, diversity and domain architecture variability in prokaryotic DNA-binding transcription factors. *PLoS ONE.* 2018;13:e0195332. doi:10.1371/journal.pone.0195332
- Ramírez-Romero MA, Masulis I, Cevallos MA, González V, Dávila G. The *Rhizobium etli*  $\sigma 70$  (SigA) factor recognizes a lax consensus promoter. *Nucleic Acids Res.* 2006;34:1470-1480. doi:10.1093/nar/gkl023.
- Romero L, Contreras-Riquelme S, Lira M, Martín AJM, Pérez-Rueda E. Homology-based reconstruction of regulatory networks for bacterial and archaeal genomes. *Front Microbiol.* 2022;13:923105. doi:10.3389/fmicb.2022.923105