

Universidade Federal de Minas Gerais

TAXI and CoryneRegNet 7.0, A creation and update of data warehouses for study speciation and regulation of transcription of prokaryotic organisms

Belo Horizonte
March 2016

Universidade Federal de Minas Gerais

**TAXI and CoryneRegNet 7.0, A creation and update of
data warehouses for study speciation and regulation
of transcription of prokaryotic organisms**

Author:

Lucas Martins Ferreira

Supervisor

J. Miguel Ortega

Co-Supervisor

Jan Baumbach

Belo Horizonte
March 2016

Abstract

TAXI and CoryneRegNet 7 - Bioinformatics platforms for analyzing prokaryotic speciation and transcriptional regulation

The evolution of informatics through time opened possibilities to other fields that were never experienced before. With biology, it was not different; it opened a new branch and gathered efforts to develop knowledge and create a new field: bioinformatics.

With sequencing projects, it generated a huge amount of data about organisms. When those organisms were compared, it was possible to notice shared genes between them, some that were more basal and others that were more specific to certain species or strain.

In prokaryotes, genes are organized in units, which are transcribed together, sharing initiators and terminators. These units can be formed by one or more genes, monocistronic and polycistronic respectively, and the construction of the unit might happen in the evolution of the organism as a whole or, in special cases, it might be specific to certain genes, species, or strains.

The control of transcription of those units is made by some special proteins called “transcription factors” and “sigma factors,” creating a gene regulatory network, another field studied by bioinformatics to understand the survival and growth of organisms.

This work compiles the efforts made to develop two systems with distinct main goals but with similarities in the developed studies. The evolution and construction of organisms in speciation could explain the creation and growth of gene regulatory networks, since the speciation process adds new features to organisms making the gene regulatory network grow.

Keywords: database; speciation; regulatory networks; transcription.

Dansk Resume

TAXI og COryneRegNet 7.0. Opstilling og opdatering af Databank for studie af speciering og regulering af transskription af prokariotiske organismer.

Udvikling af IT gennem tiderne, har givet andre studiefelter muligheder som aldrig før har været prøvede. Angående Biologien har det ikke været anderledes, nye veje er blevet åbnet hvor samlede indsatser har udviklet lærdommen og et helt nyt felt er blevet til: Bioinformatikken.

Sekvenseringsprojekter har produceret en uoverskuelig mængde af oplysninger angående organismene. Ved sammenligning af disse kan det ses at der er samme slags gener som går igen, nogle er mere basale og andre mere specifikke med hensyn til slægt eller stamme.

I prokariotiske organismer er generne organiserede i transskriptionsenheder, disse enheder er fuldstændig transskriptionerede, som svar på en initiator og en terminator, disse enheder kan bestå af en eller flere gener, således at de enten er monocistroniske eller policistroniske. Formation af disse enheder kan hænde under hele specieringsprocessen eller med eksklusivitet af visse "clados", om disse så er ældre eller nyere.

Kontrollen af transskription af disse enheder sker med specielle proteiner, som bliver kaldt transskriptionsfaktorer og sigma faktorer som generer et regulerende net af gener, dette er så et andet felt som er dækket af Bioinformatikken for en bedre forståelse af organismernes overlevelse og vækst.

I dette arbejde forenes kræfterne med det mål at udvikle to databaser med lighed af de udviklede studier. Evolution og konformation af organismene ifølge speciation kan forklare de regulerende nets tilkommen og vækst, med start på summering af nye særpræg til organismene, med vækst af de to genregulerende net.

Nøgleord: Databank, speciation, regulerende net, transskription.

Resumo

TAXI e COryneRegNet 7.0 Criação e atualização de um banco de dados para o estudo da especiação e regulação de transcrição de organismos procarióticos.

A evolução da informática através do tempo deu para outros campos possibilidades nunca antes experimentadas. Com a biologia não foi diferente, abrindo novos caminhos congregando esforços para o desenvolvimento de conhecimento e criação do novo campo, a bioinformática.

Projetos de sequenciamento geraram uma quantidade inimaginável de informações sobre os organismos. Comparando estes é observável o compartilhamento de genes através deles, sendo alguns mais basais e outros mais específicos para uma espécie ou cepa.

Em procariotos genes são organizados em unidades de transcrição, estas unidades são transcritas completas respondendo a um iniciador e terminador de transcrição, estas unidades podem ser formadas por um ou mais genes, sendo respectivamente monocistronicas ou policistronicas. A formação destas unidades pode ser durante todo o processo de especiação ou exclusivo a determinados clados, sendo eles antigos ou recentes.

O controle da transcrição destas unidades é realizado por proteínas especiais denominadas fatores de transcrição e fatores sigma gerando uma rede regulatória de genes, este sendo outro campo abordado pela bioinformática para um melhor entendimento da sobrevivência e crescimento dos organismos.

Neste trabalho são compilados esforços com o objetivo de desenvolver duas bases de dados, com semelhanças em seus estudos desenvolvidos. A evolução e conformação de organismos através da especiação pode explicar a criação e crescimento de redes regulatórias, partindo da adição de novas características aos organismos crescendo as duas redes regulatórias de genes.

Palavras chave: banco de dados; especiação; redes regulatórias; transcrição.

Acknowledgements

This PhD thesis could not have been written without the support of many people. Firstly, I would like to thank my supervisor Prof. Dr. J. Miguel Ortega and Co-Supervisor Prof. Dr. Jan Baumbach; without their support and discussion this work would never be completed.

I would also like to express my thanks to Prof. Dr. Richard Röttger with support and discussions about CoryneRegNet.

I would like to thank my colleagues from *Laboratório de Biodados*: Verônica Melo, Kátia Lopes, Tetsu Sakamoto, Ricardo Vialle, Henrique de Assis, Marcele Laux, Diego Trindade, Fernanda Stussi, and Carlos Alberto Gonçalves.

Also, my friends from the Computational Biology Laboratory: Anne-Christin, Diogo, Paulo, Nicolas, Christian, Markus, Anders, Richa and Eudes.

A special thanks to my family, Jose Isaias, Josefina Ferreira, Maria de Jesus, Miguel Noleto, Helena Bracarense, Lycianne Ferreira, João Davi, and Pedro Henrique. To my brother-in-law, Paulo Bracarense, for his support ideas and lots of beers together.

To my dear friends Thiago Rezende, Cristiano Vicente, Joaquim Paulo, Leandro Clementoni, Alexandrina Leite, Juliano Arnosti, Juliana Tibães, and Roberto Raittz.

And, the most special one, to my sister Maria Emilia for all her support, discussions, ideas, and advices.

Preface

This work was developed as a thesis to be submitted in fulfillment of the requirements for the degree of Ph.D. at the University of Southern Denmark. The work was developed by the candidate Lucas Martins Ferreira. The candidate holds a B.Sc. degree on information technology by *UniCeuma* (Universitary Center of Maranhão), followed by a M.Sc. on Bioinformatics by the Federal University of Paraná.

The work was developed on a double degree agreement with the Federal University of Minas Gerais. The work started at the *Laboratório de Biodados* in April/2012, and in June/2014 continued at the University of Southern Denmark, in the Computational Biology Laboratory.

The work started under the supervision of Prof. Dr. José Miguel Ortega, at the Federal University of Minas Gerais, and in 2014, Prof Dr. Jan Baumbach, under the double degree agreement, completed the double work supervision at the University of Southern Denmark.

The work was developed to fulfill the needs of both laboratories in the study of organisms of interest. Both projects have the main goal of study the development of bacteria and related life forms, to understand the evolution of these organisms and how they survive in response to different environments. The union of the two databases used in this work creates the basis for elucidating the evolution of organisms alongside the acquisition of mechanisms to control the transcription of acquired new genes.

List of publications

FERREIRA, L. M.; ORTEGA, J. M.; How old are the genes in *E. coli* operons? – 8th X-Meting 2012 (October 14-17, 2012) – Brazilian conference on bioinformatics. University of Campinas (UNICAMP) Campinas, São Paulo, Brazil.

FERREIRA, L. M.; ORTEGA, J. M.; THERE IS CODON USAGE BIAS AMONGST *E. coli* GENES – 9th X-Meting 2013 (November 03-06, 2013) – Brazilian conference on bioinformatics. Federal University of Pernambuco (UFPE) Recife, Pernambuco, Brazil.

FERREIRA, L. M.; RÖTTGER, R.; BAUMBACH, J.; CoryneRegNet v7.0 – Updated backend for keeping up with a growing amount data – ECCB/ISMB 2015 – 14th European Conference on Computational Biology/23rd Intelligent System for Molecular Biology. (July 10-14, 2015) Dublin Ireland.

List of Figures

Figure 1: Conformation of RNA polymerase, 1.1 A – Presents the apoenzyme of RNA polymerase, the first four fixed subunits of RNA polymerase without the sigma factor, 1.1 B – A sigma factor subunit, free on the cytoplasm and can be recruited by the apoenzyme to form a holoenzyme, 1.1 C – Formed holoenzyme with all five subunits with the ability to start a prokaryotic transcription [26].	25
Figure 2: Transcription phases: from binding site recognition to termination [42].	26
Figure 3: Binding sites positions for the sigma 70 family and sigma 54 family on the bacterial genome of Escherichia coli k12 [55] [54].	28
Figure 4: OperonDB’s interface, available at http://operondb.cbcb.umd.edu/cgi-bin/operondb/operons.cgi , where a user can perform queries or download operon predictions for more than 500 prokaryotic organisms.	35
Figure 5: DOOR database, available at http://csbl.bmb.uga.edu/DOOR/ . It is a database with computational operon predictions for more than 2,070 prokaryotic organisms.	36
Figure 6: ODB database, publically available at http://operondb.jp/ . A database for operon prediction, which combines putative operons with literature-based information.	37
Figure 7: ProOpBD: A database for operon predictions with putative operons for more than 1,200 prokaryotic genomes.	38
Figure 8: Figure displaying the system structure of the Taxonomy innovations system. The architecture is divided in two parts: a back-end database that contains all results of parting the MicrobesOnLine database, also with the result of determination of the Lowest Common Ancestor – LCA presentation on 8B; and the front-end presentation on part 8A, which presents the interconnection between the actors used to create TAXI’s web interface.	45
Figure 9: A diagram of TAXI database’s entity relation, showing all tables developed to store all biological concepts, which belong to the database scoop. The database was divided into two groups of tables. The major tables – with biological concepts and accessory tables –store relationships between the biological concepts and the features of the concepts.	46
Figure 10: TAXI’s (Taxonomy Innovations) main page, which presents the welcome image and the current version of the web page, with the navigation menu of the system.	48
Figure 11: An example of search made on the database, in the 11A part, using the option to search for an organism with the term “ <i>subtilis</i> ” in the name of the organism. The 11B part shows the query result, with all 4 organisms found on the database with the respective navigation link to the organism.	49
Figure 12: Display the sections of the organisms’ web page. Section 12A presents the mains statistic; 12B, Taxi innovations; 12C, Graphics; and 12D, genes that belong to the organism.	50
Figure 13: Graphics for organisms. The Graphics are automatically generated with MySQL, the queries were performed via PHP scripts and drawn using the Chart.js JavaScript library, publically available at www.chartjs.org . Graphic 13A shows the quantity of transcription units per size of transcription units. Part 13C shows the quantity of genes per speciation clade, 13B presents a graphic generated for the minimal LCA per TU, and 13D shows the maximum LCA per TU.	51
Figure 14: This screenshot presents the page for gene concept, with all information stored in the database for genes. The web page contains the basic information for each gene and links for transcription units and orthologue groups.	51
Figure 15: Displays the page of the transcription unit concept, divided into two parts. The first part shows the main information from TU and the second part is more related to the genes that form the operon, displaying links for the genes and orthologue groups with LCA for each gene and the position inside the TU.	52
Figure 16: This figure presents the page for the orthologue group concept with the three main parts. The first one presents basic information of the orthologue group, the second part presents the graphic of size of transcription units per quantity of transcription units, and the last part presents the description of genes inserted in the gene cluster.	53
Figure 17: Page for querying innovations on bacterial genome of a specific organism. There are three types of queries that can be performed. Query over innovations on transcription units,	

queries of innovations on genes restricted to speciation clades, and queries for transcription units restricted to a speciation clade.....	54
Figure 18: Presents the browse page, where a query can be performed for a desired organism and to access its complete statistics. The browse page is divided into two sections: a basic statistics of all genomes inserted in the database and a second section for choosing an organism for analysis.	55
Figure 19: Figure graphically presenting all transcription units of table 7 with a key linking the colors on the graphic with the clades of speciation divided by organism.	64
Figure 20: Graphic comparing the size of transcription units per quantity of calculated transcription units, considering all organisms inserted in taxonomy innovations database.....	66
Figure 21: Comparison of transcription unit size of fifteen analyzed organisms covering a vast quantity of taxons, presenting the similarity and differentiation of related and unrelated organisms. Each curve presents an analysis for one specific organism, following the same pattern observed on figure 2.13 of the transcription unit size distribution.....	67
Figure 22: Figure presenting the correlation between transcription unit sizes per quantity of transcription units with the same size. It is the same analysis presented on graphic 2.18, but only for <i>bacillus</i>	68
Figure 23: Analyses performed for other group with five bacteria, showing the curves for transcription unit sizes per quantity of TU. This figure presents a similarity between the curves of bacterial genomes.	68
Figure 24: Figure presenting the curves of transcription unit distribution per transcription unit size for <i>archaea</i> . Five out of six organisms presented on the graphic show almost the same curve conformation, except for the <i>archaea Ferroglobus placidus</i> DSM 10642.	69
Figure 25: Comparison between <i>Archaea</i> and <i>Bacteria</i> , showing a difference of the transcription units' conformation patterns of <i>archaea Halobacterium</i> sp NRC-1, which show a different pattern that is shared with other archaea, except for the archaea <i>Ferroglobus placidus</i> DSM 10642, which shows a curve similar to the <i>bacteria</i>	70
Figure 26: Comparison between all <i>Bacillus</i> bacterial genomes, amongst every clade of the bacterial speciation. The quantity of genes shared in each clade is compared.....	71
Figure 27: The figure shows the speciation of five bacteria among the clades, showing the gene acquisition for each clade. It presents the similarities and differences between different species, varying the length of the speciation process and the amount of genes acquired.....	72
Figure 28: Graphics comparing the speciation of six <i>Archaea</i> , showing different peaks of gene acquisition in comparison with the patterns presented by previous discussed bacteria.	74
Figure 29: Transcription unit conformation for <i>Bacillus</i> organisms shown in Table 3. The lines for minimum LCA represent the quantity of transcription unit conformations started per clade and the maximum LCA lines show the quantity of transcription unit conformations ended per clade, i.e., the ancestry for the first and the last genes added to the operon, respectively.	76
Figure 30: Analyses of transcription unit conformations for five bacterial genomes divided into <i>Proteobacteria</i> and <i>Actinobacteria</i> families, comparing the process through the clades of the organisms.	77
Figure 31: Figure displaying the analyses for <i>archaea</i> . The analyses were performed using the same process as for bacteria, as discussed previously. With less genes, the speciation process was influenced and it presented a different pattern for the conformation of transcription units.	79
Figure 32: RegulonDB's website, developed and published by the Centro de Ciencias Genómicas (CCG – Genomic Science Center – in Spanish) of the Universidad Nacional Autónoma de México (UNAM – National Autonomous University of Mexico – in Spanish). A reference database of bacterial genome of <i>Escherichia coli</i> k12 mg1655. The database stores biological information about operons, gene regulatory networks, and predicted binding sites. .	33
Figure 33: RegulonDB database's entity relationship.....	33
Figure 34: MicrobesOnLine's web interface. On the web page, the user can access biological information stored in the database, perform analyses, submit new data, store incomplete genomes, and gain access to the DMBS.....	34
Figure 35: PRODORIC: a database developed and published by the Bioinformatics Competence Center of Braunschweig and the Institute of Microbiology of the Technical University of	

Braunschweig. PRODORIC is an acronym for Prokaryotic database of gene regulation, and it describes a large number of controlled vocabularies to annotated information on the regulation of gene expression in prokaryotes.	84
Figure 36: PRODORIC database’s entity relation diagram, available for download on the database’s web interface, where the user can integrate on their own program, performing local queries.	84
Figure 37: Database developed and published by the Human Genome Center of the Institute of Medical Science, University of Tokyo. It is a reference for <i>Bacillus subtilis</i> studies.	85
Figure 38: MtbRegList’s website: a database developed and published by the Département de Biologie (Biology department – in French) of the Université de Sherbrooke (University of Sherbrooke – in French). This database focuses on human pathogen Mycobacterium tuberculosis, analyzing gene expressions and regulation data.	86
Figure 39: RegTransBase: is a database developed and published by the laboratories Genomics Division, Lawrence Berkeley National Laboratory, The Virtual Institute of Microbial Stress and Survival, and the Research and Training Center on Bioinformatics. The project is supported by the US Department of Energy Genomics GTL and by the Howard Hughes Medical Institute. The platform is open-accessed with a user-friendly interface; its main goal is to cover a wide microbial diversity and provide a collection of experimental data.	88
Figure 40: PePPER: a web service developed and published by the Department of Molecular Genetics of the University of Groningen. Figure showing the prediction tool of the Transcription Factor Binding Site.	90
Figure 41: Tractor DB’s web interface. A computational prediction database of TFBS and regulatory networks based on the Escherichia coli regulatory network.	91
Figure 42: SwissREGULON: A database for genome-wide annotations of regulatory sites in the intergenic regions of genomes.	92
Figure 43: Ontology-based data structure of CoryneRegNet’s entity relation diagram. The structure is divided into two parts: Ontology-based Data Structure and Generalized Data Structure. The first part is the main part of the database, storing the main tables with biological concepts. The second part houses attached information to the biological concepts and relations.	99
Figure 44: New data structure for the CoryneRegNet database. Ontology-based tables were created for biological concepts, decreasing the amount of rows per table and the time required to perform searches on the tables. Tables for organisms, genes, transcription units, and regulation units were created. This last concept was divided in regulations of transcription factors and sigma factors.	101
Figure 45: CoryneRegNet 7.0’s intro page. On the intro page, the user can find the menu to navigate on the entire CoryneRegNet’s web interface, performing searches and analyses of genes, proteins, transcription units, or gene regulatory networks. On the intro page, the user has access to documentation and download links of CoryneRegNet’s back-end software and database, and to the last database update.	106
Figure 46: A detailed CoryneRegNet’s screenshot, showing the major sections of the front-end in which the user can perform analyses and gain access to biological information stored in the database. The user can search for terms in the database, browse through genomes stored in the database, access statistics, and compare it between organisms, predict binding sites on upstream sequences of genes, and verify contradictions on microarray data.	107
Figure 47: Search page in which the user can perform searches of biological terms in the CoryneRegNet database, searching for a specific organism or for all organisms, for a special type or all types, and sorting the result by a specific type of term.	108
Figure 48: Result of a search in the CoryneRegNet database: a search performed using term “cg0444” on all organisms, in any field, and sorted by gene identification.	108
Figure 49: Statistics page in CoryneRegNet’s interface, divided into two parts: 49A, page of main statistics, general information about the database and links for pages of organisms for each organism, and part 49B, which shows further specific statistics of the regulatory networks stored in the database.	110
Figure 50: Transcription factor biding scan. The CoryneRegNet’s user can perform binding site predictions using HMMer profiles stored on the database or create their own HMMer profile	

and run the prediction on the genome sequences in the database.	111
Figure 51: Result of binding predictions executed by inserting three sequences, with the validated binding motif of the “cg0444” gene inside the three sequences. The nhmmer tool from HMMer package was able to find the correct binding site inside all inserted sequences.	112
Figure 52: Screenshot of the result of the binding motif prediction shown on figure 3.19B. A DNA sequence repeated ten times was used to generate the HMMer profile to search the motif on the complete genome sequence of <i>Corynebacterium glutamicum</i> ATCC 13032.	112
Figure 53: Contradictions on microarrays. A screenshot showing the input options of contradictions on microarrays, presenting the three ways of analyzing the contradictions in the context of gene regulatory networks stored in the database.	113
Figure 54: Contradictions on microarray results page, showing comparisons of regulatory network information with microarray data. The results were achieved using a toy test present on table q0, in which five contradictions were found, considering two auto-regulations on the calculation. This feature of CoryneRegNet could provide a background for operon predictions, elucidating errors on the microarray data or missing regulatory interactions.	115
Figure 55: Screenshot of the first section of the organism web page for the bacteria <i>Corynebacterium glutamicum</i> ATCC 13032. In this section, the main statistics of the organism are presented as quantity of genes, proteins, transcription units, regulations, and three graphics of nucleotide contents.	116
Figure 56: Second section of the organism web page. This section is optional and not shown for all organisms. Here, one organism can present modules, stimulons, both, or none.	117
Figure 57: Last section of the organism page, showing all genes for that organism with links for gene and operon pages. Also, at the bottom of the page, there is a link for the graph of the entire gene regulatory network for the referred organism.	118
Figure 58: The first section of gene concept web page containing basic information for the gene, encoded protein of the organism, and optional stimulon information that is presented if the gene is up-stimulated or down-stimulated.	119
Figure 59: Attribute section of the gene concept page that presents information such as the first and last base pair, strand, and the HMMer logo, if the gene is a transcription factor.	120
Figure 60: Screenshot of an example of the regulation section of the gene concept page. This example is based on the regulations of the gene “cg0444” of the bacteria <i>Corynebacterium glutamicum</i> ATCC 13032. This gene encodes a transcription factor with 53 regulations validated in laboratory.	121
Figure 61: Types of binding site predictions presented on the gene concept web page, where binding sites of transcription factors can be found on the upstream sequence of the current gene or if the gene encodes a transcription factor; the binding site for other genes can also be predicted.	121
Figure 62: The last section of the gene concept page with candidates for homologues of the current gene, sequences of gene and amino acid, and a link for the graph in which the current gene is inserted.	122
Figure 63: A small graph generated for gene “cg0444” of <i>Corynebacterium glutamicum</i> ATCC 13032 with a depth cut-off of one, connecting only the first layers of the regulations with the gene. On the graph, the red nodes represent the genes and the colored lines represent the regulations that connect the genes or the genes inside operons.	123
Figure 64: This screenshot presents the complete regulatory network of <i>Corynebacterium glutamicum</i> ATCC 13032 with all five sigma factors that regulates the transcription of genes in the bacterial organism.	125
Figure 65: The screenshot shows all options available for simulating the control of sigma factors on the graph generated for gene “cg0012” of <i>Corynebacterium glutamicum</i> ATCC 13032. Figure 65A shows both options activated, simulating the presence and activation of the sigma factor “cg2092 - sigA” in the graph. Figure 65B simulates the absence of the sigma factor “cg2092 - sigA” in the graph, hiding the presence of regulations for that sigma factor. The last figure, 65C, shows the presence of the sigma factor “cg2092 - sigA;” but in this example, the regulation is deactivated graining the regulation targeting to the target genes of the sigma factor.	126
Figure 66: The screenshot of the modal box for the gene “cg0012”. The modal box contains	

information and links of genes, proteins, and transcription units alongside mini graphs for validated and predicted regulations in which the gene is inserted.	127
Figure 67: Evolution of the CoryneRegNet system of all versions, from 2006-2011. The major improvements per version are presented, such as the addition of new organisms with important features, such as COMA and gene clusters.	130
Figure 68: Graphic showing the evolution of quantities of transcription factors, regulated genes, regulations and biding motifs. The graphic shows the data growth of CoryneRegNet's versions.	130
Figure 69: Graphic representing the major problem generated by CoryneRegNet's evolution, in which the exponential growth of data made the back-end unsuitable, the data structure insufficient, and the time of response high.	131
Figure 70: In the figure, the transfer of a regulation from the source organism <i>Corynebacterium glutamicum</i> ATCC 13032 to the target organism <i>Corynebacterium efficiens</i> YS-314 is seen. This transfer fits in the first type of transfer, in which the target gene of the organism that receives the regulation is a monocistronic transcription unit.	137
Figure 71: This figure represents an example of the second type of transfer covered by the new transfer pipeline, in which the target gene of the recipient organism is the first gene of a polycistronic transcription unit.	138
Figure 72: The transfer of a regulation from the organism <i>Corynebacterium glutamicum</i> ATCC 13032 to the organism <i>Corynebacterium efficiens</i> YS-314. The figure shows two arrows that indicate the presence of a main regulation and a side regulation, in which the orthologue target gene of the original organism is not the first gene of the transcription unit.....	139
Figure 73: Figure representing the regulation of the gene cg2445. It presents each regulation related to the gene, not only of transcription factors, but also the regulation by sigma factors.	144
Figure 74: Figure presenting the entire regulatory network for the organism <i>Corynebacterium glutamicum</i> ATCC 13032, with 2,800 regulations. The complete regulatory network involves almost all genes of the bacterial genome.....	145
Figure 75: Figure presenting the full-gene regulatory network for the bacterial organism <i>Escherichia coli</i> K12 MG1655, with regulations controlled by transcription factors and sigma factors.....	147
Figure 76: Figure showing the predicted regulation of the predicted transcription factor cpfrc_1525 repressing the transcription of cpfrc_1290.....	149

List of Tables

Table 1: Comparing the features of the main data analysis of related platforms.	39
Table 2: The table shows the amount of organisms used in the taxonomy innovations database, with the total of genes transcription units and orthologue groups.....	56
Table 3: Table detailing quantities of genes, transcription units, and orthologue groups of a small set of organisms inserted in the TAXI database.	57
Table 4: Table presenting the quantities of genes divided per clade on the genome speciation of all genomes studied by the taxonomy innovations system.....	58
Table 5: Table presenting quantities of genes divided per clade over the genome speciation for all the previous mentioned organisms.	59
Table 6: Table showing a list of later-acquired genes, with general information about the gene.	61
Table 7: Table showing transcription units used for analyses of conformation, studying in which speciation clade the first gene and last gene were added for the conformation.	63
Table 8: Summary and analysis of the features presented by all systems considered here.....	94
Table 9: Sequences used for the biding prediction of the figure 3.19A, predicting biding motifs in entered sequences with the insertion of the binding site “AATACTTTGCAAA” of gene the “cg0444.”	111
Table 10: Microarray data used for a toy test on the contradictions on microarrays of CoryneRegNet’s interface.	114
Table 11: Colors and line formats used to create the connection between genes on CoryneRegNet’s graphs.	124
Table 12: Table with the amount of information added to the CoryneRegNet database per version, from the first version, with only one organism, to version 6.0, with 12 organisms. ...	130
Table 13: Amount of data stored in the CoryneRegNet database in version 7.0.	131
Table 14: Comparison of response times of the same queries performed in the CoryneRegNet databases versions 6.0 and 7.0.	132
Table 15: Relation of organisms with the respective amounts of genes, transcription factors, target genes, regulations, and biding motifs.....	134
Table 16: An example of a validated regulation complemented with predicted regulations. ...	135
Table 17: An example of a predicted regulation complemented with other predicted regulations.	135
Table 18: The relation of organisms with the respective amounts of genes, predicted transcription factors, predicted target genes, predicted regulations, and predicted biding motifs.	140
Table 19: Presents the quantity of regulations for the organism <i>Corynebacterium glutamicum</i> ATCC 13032 in both levels of the database, the experimental data, a biologically validated data, and a second level with predicted information of CoryneRegNet’s back-end pipeline.	142
Table 20: Presents the quantity of regulations for the organism <i>Corynebacterium glutamicum</i> ATCC 13032 in both levels of the database, the validated data, and a second level with predicted information from CoryneRegNet’s back-end pipeline.	142
Table 21: Regulations transferred from other organisms to <i>Corynebacterium glutamicum</i> ATCC 13032.....	143
Table 22: Regulations for <i>Escherichia coli</i> K12 MG1655 present in CoryneRegNet’s 6.0 version.	146
Table 23: Updated amount of regulations for <i>Escherichia coli</i> K12 MG1655 extracted from RegulonDB with the result of CoryneRegNet’s backend run for this organism.	146
Table 24: Regulations for <i>Corynebacterium pseudotuberculosis</i> FRC41 present in CoryneRegNet’s 6.0 version.	147

Table 25: Table showing the result of CoryneRegNet's back-end run predicting regulations for the organism <i>Corynebacterium pseudotuberculosis</i> FRC41.....	148
Table 26: Table presenting examples of predicted regulations for the organism <i>Corynebacterium pseudotuberculosis</i> FRC41 from different source organisms.	148

List of Abbreviations

TAXI – Taxi innovations;

CRN – CoryneRegNet;

DNA - Deoxyribonucleic acid;

RNA - Ribonucleic acid;

LCA – Lowest Common Ancestor;

TF – Transcription Factors;

SF – Sigma Factors;

BM – Binding Motif;

TFBM – Transcription Factors Binding Motif;

NCBI - National Center for Biotechnology Information;

BLAST - Basic Local Alignment Search Tool;

DB – Data - base;

TU – Transcription Unit;

SCOP Database – Structural classification of proteins;

MOG – MicrobesOnline Orthologue groups;

KO – Kegg Orthology;

UEKO – UniRef Enriched Kegg Orthology;

Summary	
List of publications	8
List of Figures	9
List of Tables.....	14
List of Abbreviations	16
1. INTRODUCTION	19
1.1. Biological Background	19
1.2. Central dogma of molecular biology	21
1.3. Operons, transcription, transcription factors, and sigma factors	22
1.3.1. Operons.....	22
1.3.2. Transcription.....	24
1.3.3. Transcription factors	26
1.3.4. Sigma factors	27
1.4. Regulatory networks.....	29
1.5. Motivation and general aims	29
2. TAXI	31
2.1. Aims	31
2.1.1. Main.....	31
2.1.2. Specific	31
2.2. Related work.....	32
2.2.1. RegulonDB	32
2.2.2. MicrobesOnLine	34
2.2.3. OperonDB.....	35
2.2.4. DOOR.....	35
2.2.5. ODB.....	36
2.2.6. ProOpDB	37
2.2.7. Summary of contents and features of the databases	38
2.2.7.1. The content of the databases	38
2.2.7.2. Data analysis features	39
2.3. Materials and Methods	40
2.3.1. Data sources.....	40
2.3.2. Data integration	40
2.3.3. System architecture.....	43
2.3.4. Data structure.....	45
2.3.5. Visualization.....	47
2.3.5.1. Web Interface.....	48
2.4. Results and Discussion	56
2.4.1. Gene classification through bacterial speciation.....	70
2.4.2. Construction of transcription units through bacterial speciation.....	75
2.5. Conclusion.....	81
3. CORYNEREGNET 7.0.....	82
3.1. Aims	82
3.1.1. Main.....	82
3.1.2. Specific	82
3.2. Related work.....	83

3.2.1.	PRODORIC	83
3.2.2.	DBTBS	85
3.2.3.	MtbRegList	85
3.2.4.	RegTransBase	87
3.2.5.	TRANSFAC	88
3.2.6.	PePPER.....	89
3.2.7.	Tractor DB.....	90
3.2.8.	SwissREGULON.....	91
3.2.9.	Summary.....	92
3.2.9.1.	The content of the databases	92
3.2.9.2.	Data analysis features	93
3.3.	Methods.....	95
3.3.1.	Data integration	95
3.3.2.	System architecture.....	95
3.3.3.	Data structure.....	97
3.3.4.	Binding site predictions	102
3.3.5.	Homology detection.....	103
3.3.6.	Network transfer pipeline	104
3.3.7.	Visualization.....	105
3.3.7.1.	User interface.....	105
3.3.7.2.	Network visualization	122
3.4.	Results and Discussion	128
3.4.1.	CoryneRegNet's evolution	128
3.4.2.	New network transfer pipeline.....	133
3.4.3.	Novel regulatory networks.....	141
3.4.3.1.	<i>Corynebacterium glutamicum</i> ATCC 13032	141
3.4.3.2.	<i>Escherichia coli</i> K12 MG1655	145
3.4.3.3.	<i>Corynebacterium pseudotuberculosis</i> FRC41.....	147
3.5.	Conclusion.....	150
4.	CONCLUSION	151
5.	OUTLOOK.....	152
7.	BIBLIOGRAPHY	154

1. INTRODUCTION

1.1. Biological Background

Bacterial life shares genes between different clades. Some genes participate on processes where they are transferred from one organism to another, which can occur in the same phylogeny or between phylogenetically-distant bacterial genomes.

Bacterial genes can also be arranged in transcription units, in which they are sequentially organized on the same DNA strand, where those genes are transcribed with shared initiator and terminator [1]. Organizing genes in such units, apparently facilitates events of horizontal transfers, in which those group of genes could be transferred as one from the source to the target organism [2] [3] [4].

Some studies report that operons of *Escherichia coli* bacteria are, in general, kept and shared with another bacterium respecting the same structure [5]. Nevertheless, it is plausible that the sequence of gene incorporation follows a pattern where catalytic units are inserted first, followed by regulatory units.

A frequent, comparative genomic analysis is the determination of the LCA, as the study performed to understand the ancestry of *rho*-like genes [6]. Using LCA it is possible to detect in which taxonomy level a gene was developed. For example, a gene may be restricted to and shared within a family, genus, species, or even subspecies or strains.

To determine the LCA for the sets of genes from a specific genome, it is indispensable to determine the orthologue groups in which the genes are inserted.

Classifying genes with the use of orthologue clusters and determining their LCA made it possible to review the minimal set of necessary genes for a genome growth and for bacterial life preservation; it also helped us to understand the path followed by the bacterial genome until the last step of speciation. The innovations obtained at each node of the phylogeny will become apparent. The temporal composition of the operons may be revealed by themselves.

This minimal set of necessary genes creates the minimal gene regulatory network for the survival and growth of the bacterial cell. However, bacteria live in different environments, where, to survive, the bacterial cell has to send or receive, via horizontal gene transfer, new triggers for internal and external signals.

A gene regulatory network is the result of interactions of regulatory proteins, DNA sequence, and regulated target genes. Regulatory proteins, the so-called transcription factors (TF), are the major key to unlock the beginning of the transcription process inside the bacteria. Topologically, the complexity of a regulatory network is the result of genes that are regulated by more than one transcription factor, and a transcription factor binds to an upstream sequence of a list of target genes [7] [8] [9].

Alongside the transcription factors, another factor also plays an important role during the process of transcription initiation. A sigma factor is a subunit of RNA polymerase that also binds to a specific DNA binding site with the transcription factor and initiates the transcription.

These DNA sequences, in which transcription factors and sigma factors bind roughly, maintain a degree of conservation. This nucleotide sequence is called transcription factor binding motif (TFBM or shortly BM).

Some computational methods have been used to perform statistical calculations on binding motifs, using conservations of nucleotides to create models to predict the existence of similar DNA sequences in upstream sequences of different set of genes from the same organism or even to use the model to predict the existence of similar binding motifs in different organisms. Hidden Markov Models (HMM) [10][11], Artificial Neural Network (ANN) [12][13][14][15][16], Support Vector Machines (SVM) [17][18][19] are examples of computational methods used for these predictions.

Advances on the field of genome sequencing provide the prospect to reveal all features of a certain organism, presenting every peculiarity in the gene regulatory network. Starting with these approaches, researchers could not only reach new levels of knowledge based on the regulatory complexity of a certain bacterial cell, but also reconstruct the global connectivity of a regulatory network to theoretically describe it and deduce the gene expression pattern of a microorganism [20].

After several studies of bacterial genome sequencing, it was deduced that organisms which survives to different conditions found in different environments tends to carry a larger number of transcription factors, to respond to different sets of signals [21] [22].

It is also important to show that, besides an abundant amount of genome sequencing studies, the quantity of revisions of gene regulatory networks is limited to a few organisms. Also, there is no clear relationship between transcription factor, binding motif,

and target gene, that could have been horizontally or vertically transferred among genomes [23].

Two studies are, by far, the best well-characterized published reviews about gene regulatory networks. The best characterized regulatory network is the gram-negative *Escherichia coli* K12 bacterium contained in RegulonDB [24]. The second best study is the gram-positive *Bacillus subtilis* bacterium contained and documented in the DataBase of transcriptional regulation in *Bacillus subtilis* (DBTBS) [25].

Extending those studies, computational methods were used to perform analysis on complete genomes and to predict gene regulatory networks in genomes with not-so-extensive studies or no studies at all. Starting with model organisms, several groups developed techniques to transfer regulations from a source organism to a certain target organism of interest or to a set of organisms.

Every method has special peculiarities and features, some differences in the back-end, different prediction methods using homologous proteins or computational intelligent approaches. Moreover, there are some interface peculiarities that are exclusive for each tool, genome browser, binding sites prediction, and prediction of regulatory networks. Such examples of tools are RegulonDB [24], MicrobesOnline [26], PRODORIC [27] and DBTBS [25].

With the evolution of computational and biological methods, the advances of the bioinformatics field are growing exponentially, generating an extensive amount of data to be analyzed and stored in computational databases.

Alongside the evolution of these methods, all interfaces created with the special purpose of storing and performing analyses over new biological data must be upgraded, generating new computational demands and methods to present information.

1.2. Central dogma of molecular biology

The central dogma of molecular biology explains the flow of biological information inside cells of all living organisms, starting with DNA, the genetic information is transcript into RNA, which, in turn, converts the information into proteins, a functional product for the cell [28]. The central dogma of molecular biology was proposed by Francis Crick in 1956 and published in 1970 [29].

The central dogma also suggests that DNA contains all information that the cell needs to survive; in other words, all transcripts synthesized by RNA polymerase are stored in the DNA sequence, regardless of which protein it is made of. Also, RNA is a messenger that carries the information to the ribosomes [28]. Furthermore, these processes are regulated.

1.3. Operons, transcription, transcription factors, and sigma factors

1.3.1. Operons

Biologic validation and computational prediction have revealed that genes in bacterial genomes tend to be transcribed accordingly with the structure of transcription units conserved over bacterial evolution [30] [31].

These structures of transcription units are known as an important family among these functionally conserved genomic units. Moreover, these units often appear conserved in multiple genomes, performing highly compartmentalized activities in biological pathways [32] [33] [34].

This gene association is also observed in eukaryotic genomes, affecting *Drosophila* morphology, or biosynthetic pathways in *Aspergillus* and *Neurospora* [35]. On the other hand, operons were considered as the rule, rather the exception, for bacterial genomes [35].

A convincing hypothesis of how gene operons were originated, maintained, and evolved on bacterial genomes should predict the composition, distribution, and abundance of these units in Bacteria and Archaea, and its shortage on eukaryotic genomes [35].

Many researchers already published biological studies addressing the hypothesis of how these operon units originated in bacterial cells; these hypothesis were divided in five classes: selfish operon model, natal model, fisher model, molarity model, and co-regulation model [2].

The selfish operon model hypothesis proposes that an operon organization is beneficial to the cluster of genes, not to the organism that hosts the cluster [2]. Accordingly with this hypothesis, gene clusters can be propagated by vertical or

horizontal transfer, but nonclustered genes just can be inherited via vertical transfer because the mechanism that promotes horizontal gene transfer is limited by the size of the DNA that is mobilized. Clustered genes are essential for a horizontal gene transfer because a single gene cannot perform a selectable function alone; it is necessary to transfer the entire operon to perform a working function [2].

The natal model hypothesis suggests that genes are clustered into operons if they are generated by *in situ* duplication or divergence, which explains the existence of some operons in eukaryotes. In prokaryotes, operons usually encode proteins, which belong to separate proteins families [36], assembling operons from unlinked genes.

The fisher model hypothesis theorizes that genes with physical proximity reduce the rates of harmful recombination events, which can lead to the destruction of operon structures. This hypothesis suggests that natural selection would favor the construction of operons, although this hypothesis only considers free combining populations, which does not occur in most bacterial lineages. In addition, there is no evidence of co-adapted alleles in prokaryotes outside of proteins that physically interact [37].

The next hypothesis, the molarity model, suggests that the formation of operons results in a beneficially high local concentration of protein products. Regardless of what this hypothesis says, the evidence of spatial segregation of gene products in bacterial cytoplasm is growing. Also, this model does not predict the co-transcription of gene clusters, the distribution of genes already in operons, and the variabilities of these relationships among bacterial cells [38] [39].

The last hypothesis, the co-regulation model, is the basal idea of operon hypothesis, that is, genes are clustered in operons that take advantage of the same transcription machinery for starting and stopping a transcription [40]. An operon formed under this hypothesis would require an immense selective pressure for the co-transcription of each gene added to the rising operon structure, requiring immediate juxtaposition and co-transcription of previous unlinked genes. Moreover, such strong selection would be provided by a single promoter that, at first, would be controlling the transcription of a single gene. Therefore, although co-regulations may provide selection for the maintenance of an operon once it is assembled, it cannot provide selection for the original assembly of the cluster [35].

On the other hand of the operon study, several works focus in understanding the structure of operons; in other words, those studies try to reveal which genes are clustered

inside a certain operon using concurrent computational methods and laboratory validations.

In the past decades, such computational methods were developed to predict operon structure and to compare their genes in the bacterial taxa. From this point onward, it is also possible to introduce the concept that, in each clade, during the speciation process, a gene or an entire operon was added to a specific bacterial genome.

Those computational methods perform their predictions based on DNA sequence features, including physically intergenic distance in base pairs [1] [41], gene cluster conservation [31], and function communality [42]. Each of these operon predictions presents unique features and accuracy.

Unfortunately, sometimes computational prediction of operons does not reflect the truth of an operon structure inside bacterial cells, which leads to a poor knowledge on operon formation. Dominating the nuances of operons would help to build the knowledge on gene regulatory networks.

1.3.2. Transcription

During the transcription process, an enzymatic system converts a segment of a DNA sequence into a RNA strand with complementary nucleotides to the template DNA segment. The transcription method is similar to the replication process, with a comparable chemical mechanism, the use of a model strand and four major phases: binding site recognition, initiation, elongation, and termination [28].

In bacteria, the transcription process generates several types of RNA strands. In addition, other small RNA (sRNA) types were synthesized on the transcription process, but the first types are the major used by the cell [28].

The transcription is performed by a DNA dependent holoenzyme, denominated RNA polymerase. This holoenzyme is formed by five subunits; the first two subunits form the catalytic center of the RNA polymerase are the β and β' (beta and beta string) subunits, and these two subunits are present during the entire transcription process. Another three subunits are present on the RNA polymerase, two are alpha subunits (α). These subunits also belong to the core of the holoenzyme, but there is no detailed information about the function of these two subunits during the transcription process. These four subunits form

the apoenzyme of RNA polymerase. This apoenzyme is not capable of initiate a transcription because, without the last subunit, the RNA polymerase is not capable of recognizing the binding site [43] [44].

The last subunit is the sigma factor (σ). This subunit is not a permanent element of the RNA polymerase and can be changed by responding to environmental signals, and it is responsible for recognizing the binding site on the upstream sequence of gene or transcription unit [43][44], as shown in Figure 1.

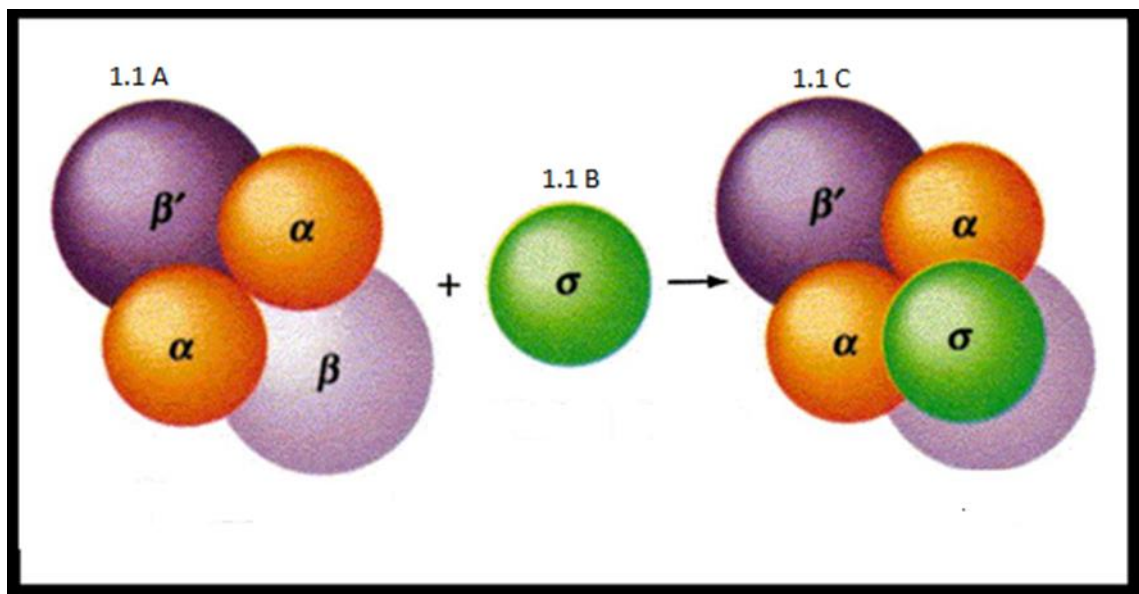


Figure 1: Structure of RNA polymerase, 1.1 A – Presents the apoenzyme of RNA polymerase, the first four fixed subunits of RNA polymerase without the sigma factor, 1.1 B – A sigma factor subunit, free on the cytoplasm and can be recruited by the apoenzyme to form a holoenzyme, 1.1 C – Formed holoenzyme with all five subunits with the ability to start a prokaryotic transcription [28].

The transcription procedure is divided into four steps: binding site recognition, initiation, elongation and termination. In binding site recognition, the transcription factors and the RNA polymerase binds to the DNA sequence segment upstream the target gene or target transcription unit for the transcription process start [43] [45].

In the initiation phase, the RNA polymerase unwind the DNA sequence creating a transcription bubble, where a limited portion of DNA is exposed, enabling the RNA polymerase to start the transcription. The second step occurs when the RNA polymerase unbinds the sigma factor and starts transcribing the DNA sequence. The third part is the elongation of the DNA sequence, where nucleotide bases are added to the RNA sequence replacing the Thymine nucleotide (T) by the Uracil nucleotide (U). The last phase of

transcription, the termination, can be performed in two ways: with or without a termination factor. Figure 2 shows all the transcription phases [44].

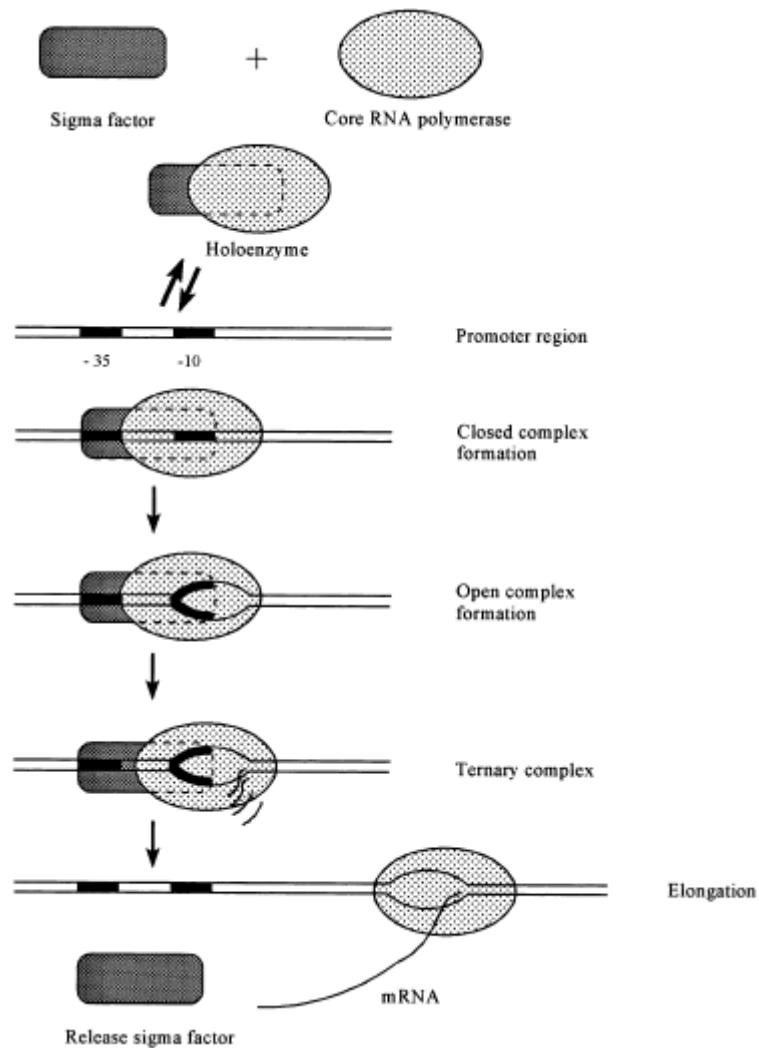


Figure 2: Transcription phases: from binding site recognition to termination [44].

1.3.3. Transcription factors

In combination with sigma factors, transcription factors play an important role during a transcription process; these proteins also bind to specific areas of a DNA sequence: binding sites and affecting the initiation of transcription, activating or repressing the transcription depending on the relative positions of their binding position to the transcription start site of a corresponding gene or transcription unit [46]. These factors bind before the RNA polymerase on a DNA upstream sequence and attract or repel the holoenzyme.

In comparison with sigma factors, transcription factors also bind to a limited set of genes controlling the transcription in response to external or internal triggers running the transcription concurrently to sigma factors [47].

Studies on *Escherichia coli* k12 genomes report the prediction of 270 transcription factors, which represents 6% of all protein-coding of bacterial genome, which, based on the hierarchical classification of SCOP database, divides these coding-proteins into 11 families [48].

Over 75% of all transcription factors predicted in *E. coli* k12 contains an additional domain belonging to a wider range of 46 different protein families, domains relating to sensing signals. Containing a second domain that can potentially bind to second modules, there is a group of 40-50% of all transcription factors [49] [50]. Another 10% of transcription factors is part of two-component signaling cascades, which are biological systems that respond to external signals in which one protein acts as a sensor that is phosphorylated by an upstream histidine kinase [51].

If the bacterial organisms vary in size, the transcription factors will vary in number. With larger and more complex genomes, an excess of transcription factors is necessary to regulate specialized groups of genes; or it may use more complex cascades of regulatory proteins [52] for that end.

The size of genomes and the complexity of gene regulatory networks agree with the environment of insertion of the bacteria. Free-living bacteria or of multiple habitats tend to have a bigger genome and a more complex regulatory network to respond to environmental variations. On the other hand, organisms that live in a symbiotic association or in parasitism have a poor TF gene content. They further emphasize the role of TFs in ensuring signal-dependent cellular responses [48].

1.3.4. Sigma factors

These subunits are not fixed on the RNA polymerase and can be altered depending on the environmental conditions where the cell is inserted. [53] [54] [55].

Some published studies on the bacterial genome of *Escherichia coli* k12 present a division of sigma factors into two families. The first family, the sigma 70 (σ_{70}) family, is related to the survival and growth of bacteria [55]. Concurrently, the second family is

the sigma 54 (σ_{54}) family, which is an alternative sigma factor family related to biological nitrogen fixation and sporulation, for example [56].

Also, these studies discovered the presence of seven sigma factors on *E. coli* k12 bacterial genome, subdivided into their families. The sigma 70 family is composed by six sigma factors: sigma 70 (σ_{70}) by itself, sigma 38 (σ_{38}), sigma 32 (σ_{32}), sigma 28 (σ_{28}), sigma 24 (σ_{24}), and sigma 19 (σ_{19}). Sigma 54 (σ_{54}) is the only sigma factor that creates an alternative sigma factor family [44] [57].

Each family binds to a specific bind site on the upstream sequence of the gene. The sigma 70 family binds to two hexamers located at the positions -35/-10 upstream of the transcription start site. For the sigma 54 family, the binding site is a little bit different; the hexamers for this family are located at -24/-12 base pairs before the transcription start site. The binding sites for both families of sigma factors of *Escherichia coli* k12 is shown in Figure 3 [57] [58] [59] [56].

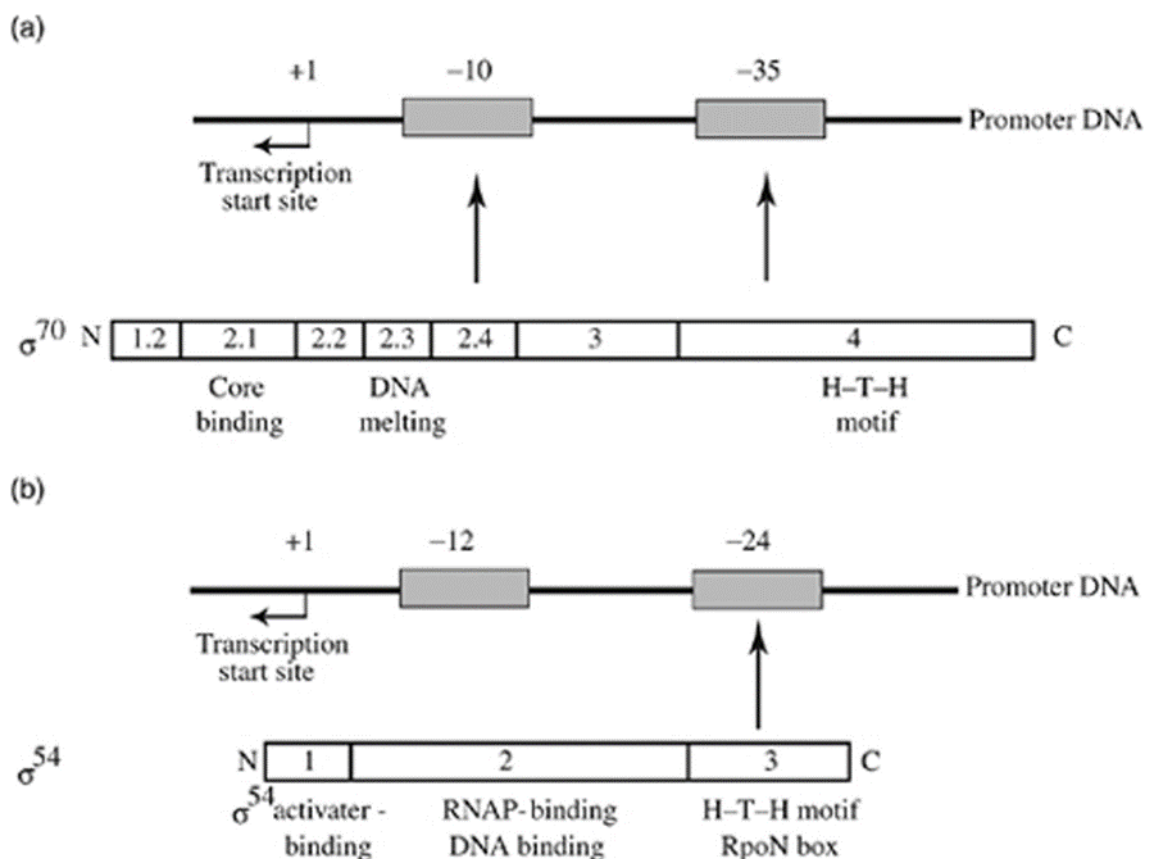


Figure 3: Binding sites positions for the sigma 70 family and sigma 54 family on the bacterial genome of *Escherichia coli* k12 [57] [56].

1.4. Regulatory networks

Over the last years, a considerable amount of information have been accumulated on regulatory interactions between transcription factors and regulated target genes in various prokaryotic organisms, such as *Escherichia coli* k12 [24] and *Bacillus subtilis* [25].

The investigation of interactions between transcription factors and regulated target genes, as a network, provides a great framework to identify principles that rule such biological systems.

In any gene cell, thousands of genes are presented at once ensuring the survival and growth of the cell. Each gene must be transcribed in proper time and amount to ensure the appropriate functional outcome [60].

These networks develop a central role for bacteria, and deciphering these networks is vital for understanding the development, functioning, and pathogenicity of these bacterial organisms [61].

The expression of basal genes is invariable. Their expression is robust and controlled by fixed regulatory networks. On the other side, some genes depend on more adjustable regulations, responding to internal or external triggers [60].

Even in closely related species with high similarity between genome sequences, gene expressions can be quite different; this divergence plays an important role in evolution of bacterial species and it is believed to be one of the primary sources of phenotypic variation between species.

Performing computational studies in the field of gene regulatory networks can prevent an unnecessary spent of resources in developing analyses in experimental laboratories or performing less expansive studies, taking advantage of computational and biotechnological evolution [62].

1.5. Motivation and general aims

Unraveling secrets hidden inside a bacterial genome leads to an understanding of how a bacteria survives or grows in different environments, responding to changes of the ecological conditions to overcome stress situations like heat shock, pH variation, or

availability of nutrients [1][2].

With the use of these internal mechanisms, bacteria can develop different molecular strategies to perform adjustments on gene regulatory networks transcribing a different sets of genes in each specific situation, using proteins capable of interacting with DNA sequences [65][66][67].

Orthologue genes have been successfully mapped into complete bacterial genomes, using comparisons of amino acid sequences. The bacterial genome of *Escherichia coli* k12 has shared genes with other bacteria and archaea, but there are exclusive genes present only on the *gammaproteobacteria* class.

Starting with these mechanisms and comparisons between organisms, some biological questions were addressed: “*How genes for a specific bacterial demand were added to the bacterial genome? How the addition of such genes influenced the creation of operons? And during the process of horizontal transfers, how these genes behave during horizontal transfer?*”

This work focuses on the study of bacterial genomes, performing the implementation of a new systems biology databases relating to taxonomy innovations, TAXI, and with the update of a well-known database, CoryneRegNet.

TAXI is a systems biology database that was developed with the use of certain techniques to determine the LCA [6] and operon predictions for more than 1,700 bacterial genomes [68]. It performs a study of gene ancestry and structure of operons on speciation, and of taxonomy innovations on bacterial speciation.

CoryneRegNet is already a well-known systems biology database of a study of *Corynebacterium* bacterial genomes, presenting complete genomes, predicted operons, predicted transcription factors binding motifs, and validated and predicted gene regulatory networks.

2. TAXI

2.1. Aims

2.1.1. Main

The main objective of the work described in this session was to analyze the clades of gene acquisition and the composition of transcription units of prokaryotic genomes in respect of the clades of acquisition of their genes.

2.1.2. Specific

Build a local database, beginning with the information provided by MicrobesOnLine database, using their transcription unit predictions and their orthologue gene clusters (MOGs).

Determine the ancestor clade, i.e. the Lowest Common Ancestor (LCA), for every gene in the local database.

For each transcription unit, to use LCA determination to obtain the minimal and maximal gene LCA in transcription units for all organisms stored in the local database.

To identify genes and transcription units related with taxonomic innovations, i.e. genes or operons that were added to a bacterial genome in one specific clade along evolution.

To develop a user-friendly web interface that helps researchers to perform analyses on taxonomic innovations along cladistics evolution.

2.2. Related work

Taxi information is unique, but there are several other databases focused on information about transcription units and they are revised here for comparison

2.2.1. RegulonDB

RegulonDB is an internationally, well-known database that is based on the knowledge of bacterial genome of the model organism *Escherichia coli* K-12 mg1655 and that publically offers information related to gene regulatory networks, activation and repression, operon organization, including their various transcription units, and the integration of regulons as sensor units. It has the major goal of compiling and editing the knowledge generated by the international scientific community, and is the major electronically-encoded regulatory network currently available for any organism [24].

All data deposited in the database is acquired by manual curation, an effort accomplished by the RegulonDB team starting with associated references obtained from PubMed, references that are also related to the evidence code that establishes distinctions between strong and weak objects [24].

The access to RegulonDB data is made by an online front-end, shown in Figure 4, where all biological information is available for querying. For one gene of interest, one can query for the available the gene product, position in the genome, molecular weight, functional classification, and the corresponding gene/protein sequence. Other features of the bacterial genome publically accessible are: a genome browser including operon organization, binding sites, promoters and terminators, and a regulatory network visualization (graph) with all known gene regulatory interactions.

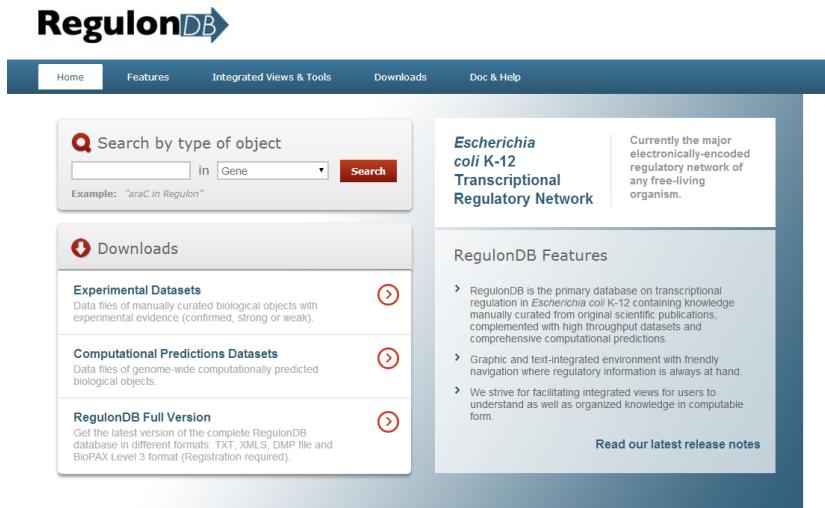


Figure 4: RegulonDB’s website, developed and published by the Centro de Ciencias Genómicas (CCG – Genomic Science Center – in Spanish) of the Universidad Nacional Autónoma de México (UNAM – National Autonomous University of Mexico – in Spanish). A reference database of bacterial genome of *Escherichia coli* k12 mg1655. The database stores biological information about operons, gene regulatory networks, and predicted binding sites.

RegulonDB web site also offers the free download of the database dump in different formats. Covering almost all DataBase management systems available on the market, it is also present in delimited flat file tabs [24]. The database structure, shown in Figure 5, covers all data available in the web site, giving the opportunity for the researcher to integrate RegulonDB with his/hers own projects.

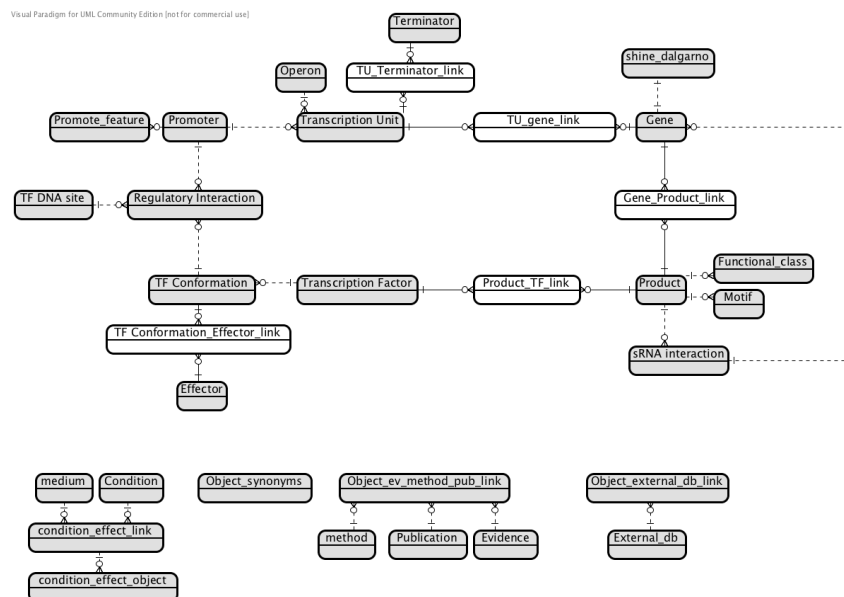


Figure 5: RegulonDB database’s entity relationship.

RegulonDB can be accessed at (<http://regulondb.ccg.unam.mx/index.jsp>) and it was developed by the Centro de Ciencias Genómicas (CCG – Genomic Science Center – in Spanish) of the Universidad Nacional Autónoma de México (UNAM – National Autonomous University of Mexico – in Spanish).

2.2.2. MicrobesOnLine

MicrobesOnLine database is the union of systems biology tools for genomic analysis in which biological data is publicly available on their front-end (www.microbesonline.org) and the access to raw data in the database is possible by direct connection with their database management system [26].

This database was developed by the Virtual Institute for Microbial Stress and Survival (VIMSS), of the United States Department of Energy. The database was part of a bigger project and it was separated from the project due to the growing amount of relevant biological data for the scientific community [26].

By accessing the database, the user can access data of genes, operon predictions, clusters of genes with similar functions, and a lot of other biological data [26]. MicrobesOnLine’s web interface is shown on Figure 6.



Figure 6: MicrobesOnLine’s web interface. On the web page, the user can access biological information stored in the database, perform analyses, submit new data, store incomplete genomes, and gain access to the DMBS.

2.2.3. OperonDB

First released in 2001, OperonDB is a database that contains the results of a computational method for operon prediction in bacterial genomes. The database started with 34 genomes in its initial release and grew to more than 500 genomes on the current version [69].

The database is publically available at (<http://operondb.cbcb.umd.edu/cgi-bin/operondb/operons.cgi>); it is constantly updated with finished prokaryotic genomes available at GenBank. All predictions can be downloaded in bulk, and the OperonDB database may be downloaded as an open source software. An image of the web interface is shown in Figure 7.

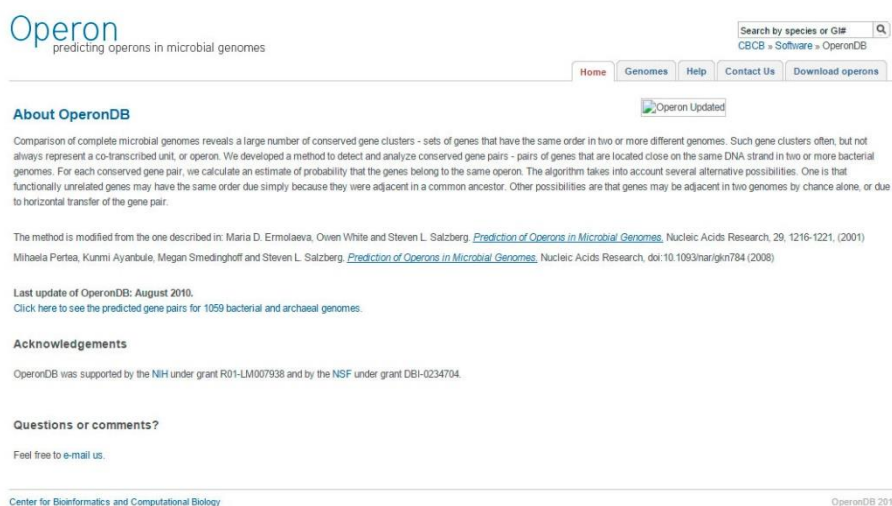


Figure 7: OperonDB's interface, available at <http://operondb.cbcb.umd.edu/cgi-bin/operondb/operons.cgi>, where a user can perform queries or download operon predictions for more than 500 prokaryotic organisms.

2.2.4. DOOR

DOOR is the acronym for Database of Prokaryotic Operons, which contains computational predictions of bacterial operons; it covers operon predictions for 2,072 organisms. DOOR also presents several queries on the database to facilitate the access for the biological information stored in it [70].

The database has a search function so the user may query for the desired operons and associated information through multiple querying methods. The database also presents a

search function so the user may find operons with similar compositions and structures [70].

Another function of the database is to search for motifs in the promoter regions of a user-specified group of possible co-regulated operons with the use of motif-finding tools. DOOR also includes database predictions for RNA genes [70].

OperonWiki is a feature available in DOOR, where the user can interact with database developers. On the interface, shown on Figure 8, DOOR database also provides links for operon predictions on other databases, linking biological data.

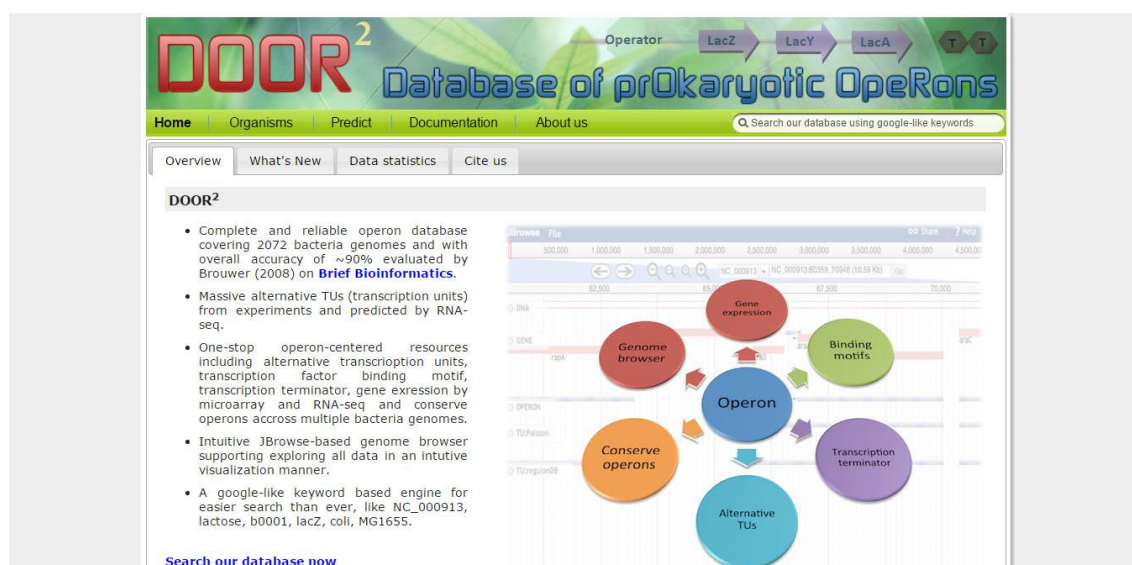


Figure 8: DOOR database, available at <http://csbl.bmb.uga.edu/DOOR/>. It is a database with computational operon predictions for more than 2,070 prokaryotic organisms.

2.2.5. ODB

Operon DataBase is a database with a data retrieval system of several numbers of already-published operons on many complete genomes. Additionally, with validated data, the database also provides predicted operons that are conserved in terms of operons [71].

The current ODB's version, ODB3, stores a total of 10,000 known operons, which belong to more than 50 bacterial genomes. Also, the database stored a total of 400,000 putative conserved operons from 1,000 genomes [72].

ODB database integrates the use of four associations: genome context; gene co-expression, obtained from microarray data; functional link on biological pathways; and

gene conservation through genomes. These associations indicate genes organized in operons and give more accuracy to operon prediction [71].

The use of computational predictions with literature-based information provides a bioinformatics tool, which can be used not only by bioinformaticians on their researches, but also by experimental biologists; the database is publically available at <http://operondb.jp/>. Figure 9 presents a screenshot of ODB's web interface.

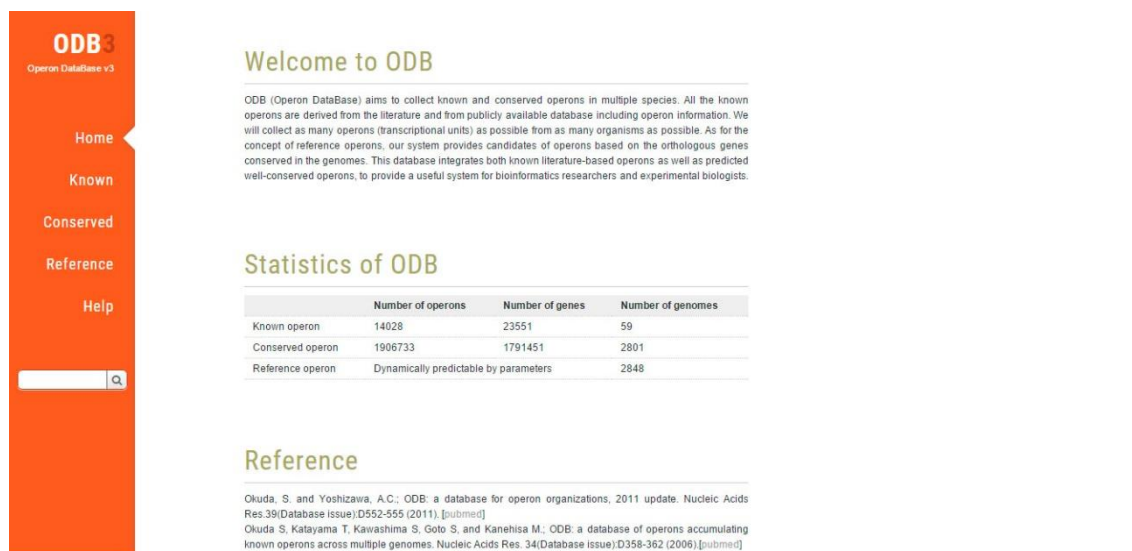


Figure 9: ODB database, publically available at <http://operondb.jp/>. A database for operon prediction, which combines putative operons with literature-based information.

2.2.6. ProOpDB

Prokaryotic Operon Database constitutes a computational repository of operon predictions. This database stores operon predictions for more than 1,200 prokaryotic genomes [73].

On ProOpDB, a set of operons can be retrieved using the name of the organism, metabolic pathways, gene orthology, conserved protein domains, reference gene, and reference operons. Moreover, ProOpDB's web interface provides a gene context tool to present the gene and its surroundings [73].

Based on an Artificial Neural Network, the operon predictor extracts some characteristics from the sequence to predict operon structures using the distance between genes and the function relationships between them [73].

The Gene Context Tool is filled with biological information provided by a series of subroutines and modules; the gene context displays the structure of predicted operons using a set of Perl-CGI programs that use the open source code GD graphics library and JavaScript codes to create the HTML files [73]. Its interface is shown in Figure 10.

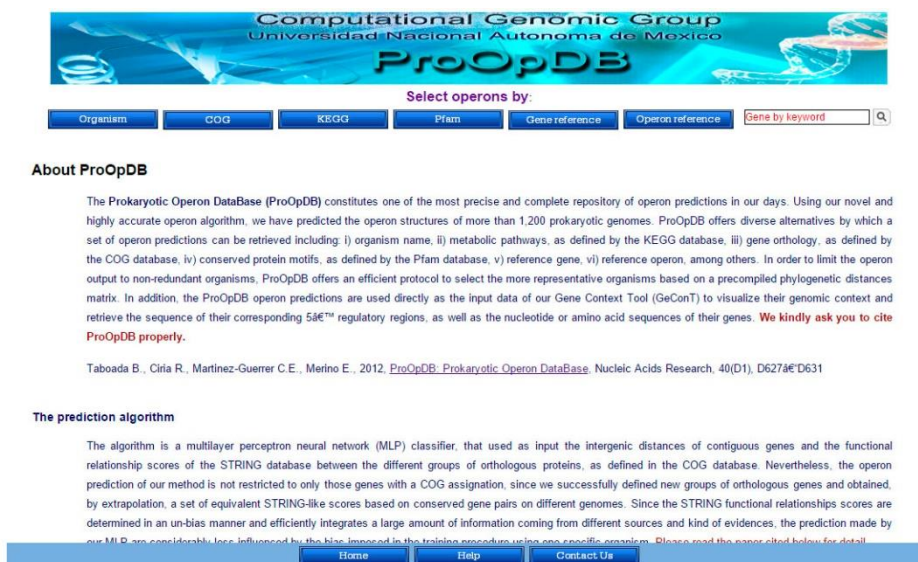


Figure 10: ProOpBD: A database for operon predictions with putative operons for more than 1,200 prokaryotic genomes.

2.2.7. Summary of contents and features of the databases

This section summarizes the mentioned related works, comparing the contents of the databases and analyzing their features.

2.2.7.1. The content of the databases

The related works store the following organisms in their databases, being mentioned here those that are interesting for this work:

- RegulonDB: 1 organism - *Escherichia coli K12 mg1655*;
- MicrobesOnLine: 3707 organisms - 1752 bacteria, 94 achaea and 119 eukaryotes;

- OperonDB: 500 organisms - *Corynebacterium aurimucosum* ATCC 700975, *Corynebacterium diphtheriae* NCTC 13129, *Corynebacterium efficiens* YS-314, and more;
- DOOR: 2072 organisms - *Corynebacterium aurimucosum* ATCC 700975, *Corynebacterium efficiens* YS-314, *Corynebacterium pseudotuberculosis* C231, and more;
- ODB: 59 known organisms, 2,081 conserved, and 2,848 dynamically predicted – *Bacillus subtilis*, *Escherichia coli*, *Salmonella enterica*, and more;
- ProOpDB: 1,200 organisms - *Corynebacterium aurimucosum* ATCC 700975, *Corynebacterium efficiens* YS-314, *Bacillus subtilis*, *Escherichia coli*, and more.

2.2.7.2. Data analysis features

In Table 1, we compare and summarize the main data analysis of related platforms setting a course for the development of our work.

Table 1: Comparing the features of the main data analysis of related platforms.

Features	RegulonDB	MicrobesOnline	OperonDB	DOOR	ODB	ProOpDB	TAXI
Operons	+	+	+	+	+	+	+
Orthologue groups		+			+		+
Operon statistics	+	+		+		+	+
LCA per gene							+
Operon LCA information							+

2.3. Materials and Methods

This section describes all materials and methods used to create the database TAXI. It started by consulting already-published databases, acquiring data to perform analysis of LCA, developing a local database to store all achieved results and, finally, implementing a web interface to present all analyses results on a user-friendly way.

2.3.1. Data sources

TAXI development required some biological background information, such as orthologue groups, operons predictions, and phylogeny; a large number of databases were already produced with results of this kind of biological information.

During TAXI development, two publically published databases were used: RegulonDB [24] and MicrobesOnLine [68]. These two databases perfectly fit the implementation needs with predicted and validated data.

The first database that was used was RegulonDB, from which the information of operon prediction for *Escherichia coli* k12 was acquired, with the aim of understanding the structure of operons of that bacteria.

To extend the studies and cover as much as possible bacterial taxa, they were introduced to the MicrobesOnLine database, gaining access to their publically published database of operon predictions and orthologue groups (named MOGs).

2.3.2. Data integration

The starting point of the TAXI database was acquiring data from RegulonDB. All information about genes and operon predictions of the bacterial genome of *Escherichia coli* k12 was analyzed and copied to a local database.

Alongside the analyzed information from RegulonDB, information of gene clustering for that bacteria were also brought together; at first, the gene clustering data from UEKO was used, a later explanation about UEKO is given on the next section.

Later, for the expansion of analyses, the data source was replaced with MicrobesOnLine, from where the operon predictions of 1,753 bacterial organisms, covering a vast amount of taxa clades, were used.

Also, MicrobesOnLine orthologue clusters replaced the use of UEKO clusters. MOG were build using analyses of a vast volume of genomes in which the genes were grouped based on gene synteny and functional similarity [68]. While UEKO clusters collect all genes that go under lateral gene transference, MOGs will present more clusters for that group of orthologues, depending on the synteny and, therefore, operon composition.

The last information added to the TAXI database was the LCA, which is calculated with the use a web service available at *Biodados* Laboratory's website (biodados.icb.ufmg.br/services). The information, which is calculated for each gene, is stored with gene clusters for a faster retrieval.

2.3.2.1. *Biodados* laboratory tools

2.3.2.1.1. UEKO

UEKO is a database developed by Gabriel Fernandes, a researcher at *Biodados* laboratory, this database is based on KEGG ORTHOLOGY annotation database, enriching it with information from UniRef50 cluster from UniProt [74].

KO database uses only genes from complete sequenced organisms to build their clusters, this lack of biological information generates a bias on the LCA determination, if a cluster is structured with genes from non-complete sequenced organisms, the result of LCA determination might be closer to the root of cladogenesis then the calculated.

The algorithm of database generation of UEKO recruits to the KO clusters sequences of genes with a percentage of similarity equal or superior of 50%, In other words, if a UEKO cluster have one gene of a UniRef50 cluster, the complete UniRef50 cluster is clumped on the UEKO cluster, since the complete UniRef50 cluster already have a coverage of 50% or more for all sequences on the cluster.

The enrichment of the database raises the number of genes in clusters in 104%, from 1.411.402 orthologue genes on the original KO database to 2.881.880 orthologue genes on the new UEKO database. With these new clusters gives, a higher accuracy was

achieved on the determination of LCA.

2.3.2.1.2. LCA determination

The LCA determination is largely used in this work; the determination is based on the *taxa* group submitted for calculation. The concept of LCA is based on the graphs theory, where the LCA determination for two nodes “*a*” and “*b*”, on a tree “*t*”, is defined as the nearest node from “*t*”, which has “*a*” and “*b*” as descendants. However, a node is considered his own descendent. This concept can be extended for the taxonomy tree on NCBI, where the LCA is the lowest common ancestor between two or more clads.

The method of LCA determination consists on the load of the entire taxonomy tree on memory, as a graph, and receiving the *tax ids* for determination, the script performs a comparison of the list of *tax ids* against the taxonomy tree, Finding the correspondent nodes for the *tax ids* list the script return the LCA for the two or more taxonomy ids queried.

For performing LCA determination, the user can use more than one method to access the script, the user can execute the determination via command line or send the list of taxonomy ids to the Web Service via SOAP or REST.

For command line use for LCA determination exists two scripts, one, *LCARunner*, for one list of Taxonomy ids and a second script, *multiLCA*, for use of multiple files for LCA determination.

For integration in systems using remote connections, the user has also two options, SOAP or REST. For SOAP access the user uses the WSDL available at <http://merengue.icb.ufmg.br:8080/BioToolsService/services/lca?wsdl>. The user sends as parameter the list of Taxonomy ids and receive an object with information for the Taxonomy ids sent. For REST use, the researcher uses the link <http://merengue.icb.ufmg.br:8080/BioToolsService/lca/txid1+txid2+txid3+%E2%80%A6>, where txid1, txid2... are the Taxonomy ids which the user wants the LCA determination.

2.3.2.2. MicrobesOnline orthologue cluster

MicrobesOnline Orthologue Groups – MOG is an orthologue group build by VMISS to identify functional orthologue groups of genes; these groups of genes only consider genes which shares same functionality, excluding from these clusters genes which participated of horizontal gene transfer process.

MOG are build starting from tree-orthologues computed by MicrobesOnline for a gene, by examining the pre-computed gene trees for that gene. This tree-orthologues have a relation of 1:1 for orthologue genes, this limitation refers to a gene in organism “A” have only one orthologue gene in organism “B”, but there are a few exceptions occasioned by inconsistencies between trees [26].

To support phylogenetic analyses of MicrobesOnline, all tree-orthologue computed for a gene were clustered generating MOG clusters, these clusters were the union of tree-orthologues [26]. These clusters creates clusters of genes more related with the functionality of the gene, being more specific than other clusters [26].

2.3.3. System architecture

The TAXI was developed as web-based software publicly available at biodados.icb.ufmg.br/taxi/, and the user can access all resources and pipeline results accessing the web interface.

Besides the web interface, there is a back-end program responsible for parsing, analyzing, and storing all biological information to give support for TAXI’s front-end. A database and a parser compose the back-end.

The TAXI parser is developed in java programming language that reads the biological information from MicrobesOnLine and converts the information to fit the TAXI database.

Basic information of bacterial organisms are copied from the online database, such as Taxonomy identification number and name of the organism to the genes list, which compose the gene orthologue cluster for each gene.

The parser starts querying the taxonomy identification number – TaxonomyID – for

each organism and the identification number is stored with the complete name of the organism; these two data are the basic information for the entire process.

All genes belonging to each organism are queried, and the query result is composed of: Locus ID (Unique identification for each genome on the database), Gene (the identification for each gene), Gene Name (name for the gene), Protein (protein name for each gene product), Strand, start and end positions relative to the first base pair, gene sequence, and protein sequence.

Afterwards, the transcription unit - TU predicted information is queried, creating the unions of genes, which were transcribed together, on the local database. For TUs only the identification of the TU is downloaded to join the genes inside the units.

The MOG cluster identifier is downloaded also. These clusters interconnect genes across different organisms, opening the opportunity to study the steps of speciation of an organism and to understand the process of construction of a transcription unit.

The last information added to the TAXI database is the determination of the LCA, generated by the querying of the aforementioned web service with the use of a SOAP requisition, sending all genes taxonomy IDs from the cluster and receiving, as a result of the calculation, the most ancestral clade in which that gene is found.

All that information can be accessed via the aforementioned web interface, where all information is concatenated and presented in a way humans can understand. The web interface was developed using HTML, JavaScript, and PHP scripts, in which each language performed one important role for the interface development. In Figure 11, the system architecture of TAXI system biology tool is displayed. For further information about the interface, see section 2.3.3 – Visualization, which contains a more detailed explanation about it.

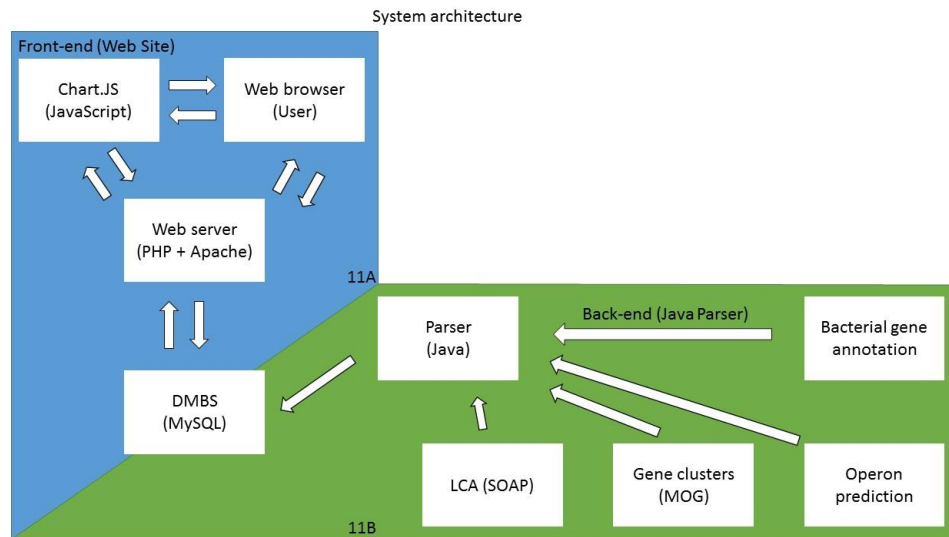


Figure 11: Figure displaying the system structure of the TAXI system. The architecture is divided in two parts: a back-end database that contains all results of parting the MicrobesOnLine database, also with the result of determination of the LCA, presented on 11B; and the front-end presentation on part 11A, which presents the interconnection between the actors used to create TAXI's web interface.

2.3.4. Data structure

The TAXI database was developed respecting all biological concepts, which were actors in the proper development of the database. All storing tables followed the central idea of the biological concepts and their relations.

The database is divided in four major tables and in another four accessory tables. The major tables house TAXI's major biological concepts: Organisms, Genes, Transcription Units, and Orthologue groups. In addition, the accessory tables store the relations between the concepts and the features of the concepts. In Figure 12, a representation of the diagram of TAXI database's entity relation is displayed.

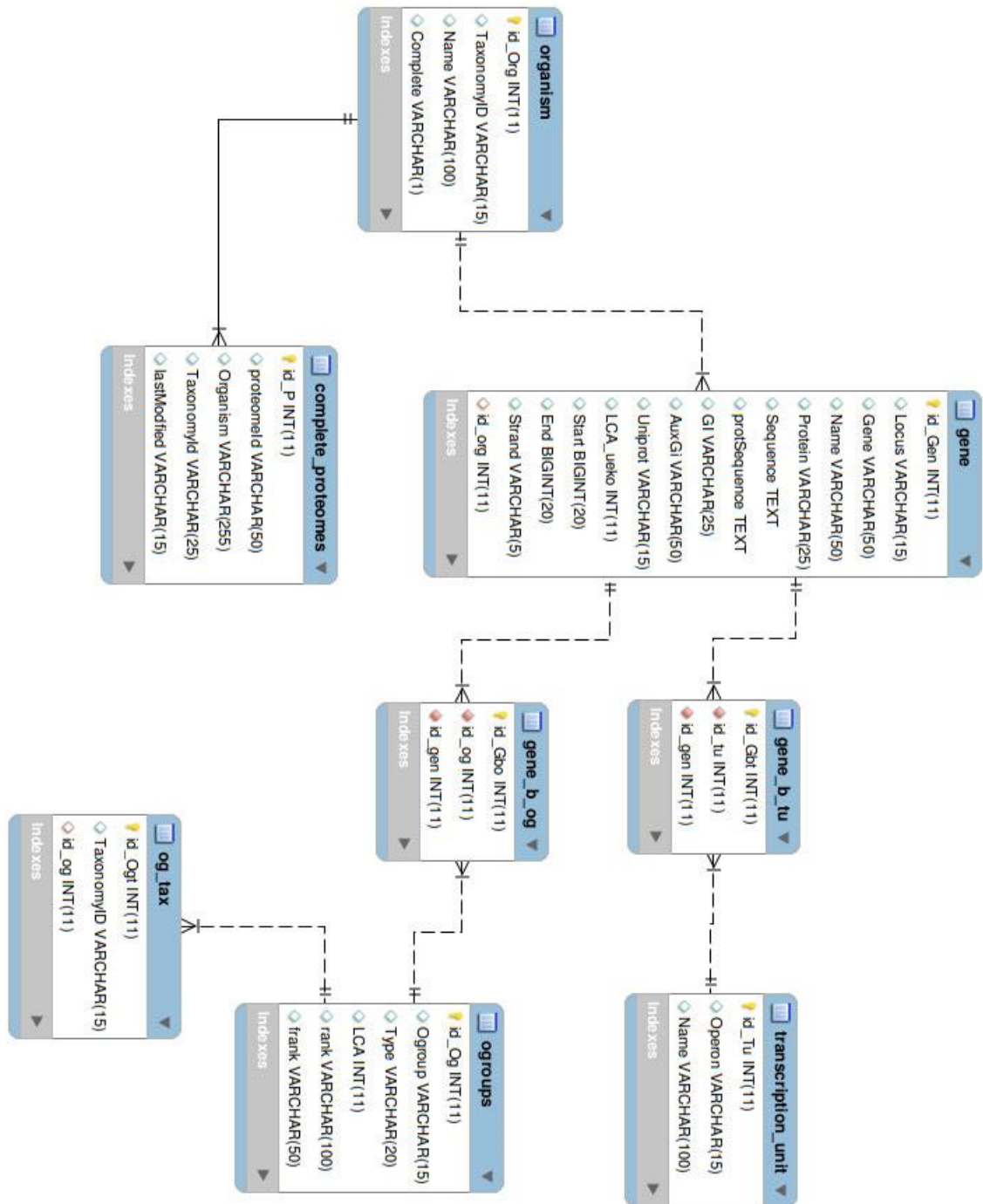


Figure 12: A diagram of TAXI database’s entity relation, showing all tables developed to store all biological concepts, which belong to the database scope. The database was divided into two groups of tables. The major tables – with biological concepts and accessory tables –store relationships between the biological concepts and the features of the concepts.

The first and most important table of the TAXI database is the organism table. This table stores the concept of the organism, and is the starting point of other tables’ analyses; this table stores the taxonomy identification number and name of the organism.

Directly connected with the organism table is the gene table, which stores the

homonym concept; this table is connected with the organism table by a foreign key of the organism ID. Also, this table stores the locus, internal unique identification of MicrobesOnLine database, Gene, gene identification, Protein, protein identification, gene name, strand, start and end positions relative to the first base pair of the organism, gene sequence, and protein sequence.

Another concept used on the TAXI database is the transcription unit. The table for this concept stores the identifications of TUs, even if there is only one gene inside the transcription unit. It also stores the identification number of the predicted TU accompanied by its name.

The last concept addressed by the TAXI database is the orthologue group. This concept table stores not only the information around the cluster, but also the LCA determined for that orthologue group. This database also stores the identification number of the orthologue group, the LCA in integer data type for comparisons across organisms (class, order, family, or no rank), rank, and first rank for LCA.

The four accessory tables for storing the relations between the concepts and the features of concepts are: the relation among gene and transcription units (where the internal identification of the transcription unit and the genes are stored) and the relation between orthologue groups and genes (a table storing only the identification of genes and orthologue groups). Also, connected to the orthologue concept table, is the taxonomy ID number table of the orthologue group, which stores only the ID number of the orthologue group with the taxonomy ID of all other taxonomy IDs of other genes, which compose the orthologue group. This last table is used only for determination of LCA. The last table of the database is the complete proteomes, which stores the information related to reference proteomes of the Uniprot consortium.

2.3.5. Visualization

This section will discuss the methodology and the final front-end of TAXI, and it will show the possibilities of the front-end introducing the navigation of the TAXI web interface.

- Query: In this section, the user can perform queries on the database about the innovations of organisms during speciation;
- Browse: This section allows the user to navigate through the organisms available on the database;
- Documentation: The user can find the entire documentation explaining the use of the database front-end and links to download the pipeline for database creation;
- Contact: Here the user finds the contact of all researchers involved on the development of the database.

On the search section, the user, as mentioned above, may perform searches on the database querying for organisms, genes, proteins, transcription units, and orthologue groups. When a user performs a search, the PHP queries the database and, if a result is found, it is presented in the result page; if no records are found, the web interface simply cleans the search fields. Figure 14 presents an image with an example of a search made on the database.

14A

Search

Perform a search

Choose a concept:

Search

Search result

14B

4 organisms found

TaxonomyID	Name
535026	Bacillus subtilis subsp. subtilis str. NCIB 3610
535025	Bacillus subtilis subsp. subtilis str. JH642
535024	Bacillus subtilis subsp. subtilis str. SMY
224308	Bacillus subtilis subsp. subtilis str. 168

Figure 14: An example of search made on the database, in the 11A part, using the option to search for an organism with the term “*subtilis*” in the name of the organism. The 11B part shows the query result, with all 4 organisms found on the database with the respective navigation link to the organism.

Using as example the bacterial genome *Bacillus subtilis* subsp. *subtilis* str 168, taxonomy ID: 224308, Figure 15 presents the web page for that organism. The page is divided in 4 sections: main statistics, graphics, taxonomy innovations, and genes. A further explanation of each section is given below.

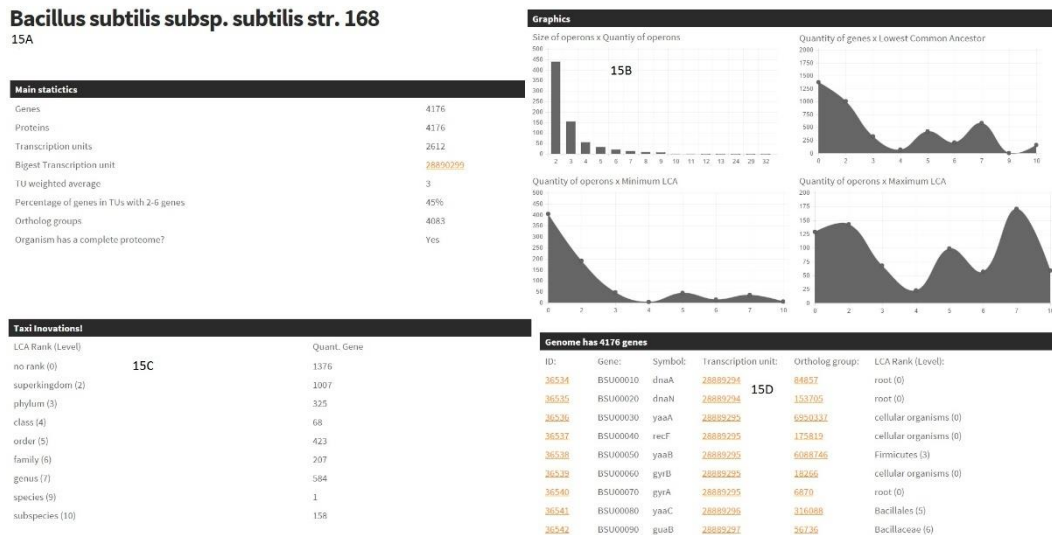


Figure 15: Display the sections of the organisms' web page. Section 15A presents the mains statistic; 15B, Taxi innovations; 15C, Graphics; and 15D, genes that belong to the organism.

Section 15A presents the main statistics of the organism: quantity of genes, proteins, and transcription units inside the organism, a link to the bigger predicted transcription unit present on the organism, TU weighted average size, percentage of genes inserted in TU with the size from 2 to 6 genes, quantity of orthologue groups inside the organism, and an optional information if the organism is a reference organism (complete genome in UniProt).

Part 15B shows the quantity of innovations that appears in each organism per speciation clade, and declares the integer used to present the clade in graphs.

15C shows all graphs generated per organism; an additional explanation about the graphs is presented on the label of Figure 16.

15D presents all genes that belong to the organism with related links to the gene information page, transcription unit page, and orthologue group page. Also, for each gene, are presented the gene name, gene symbol, and the LCA for that gene.

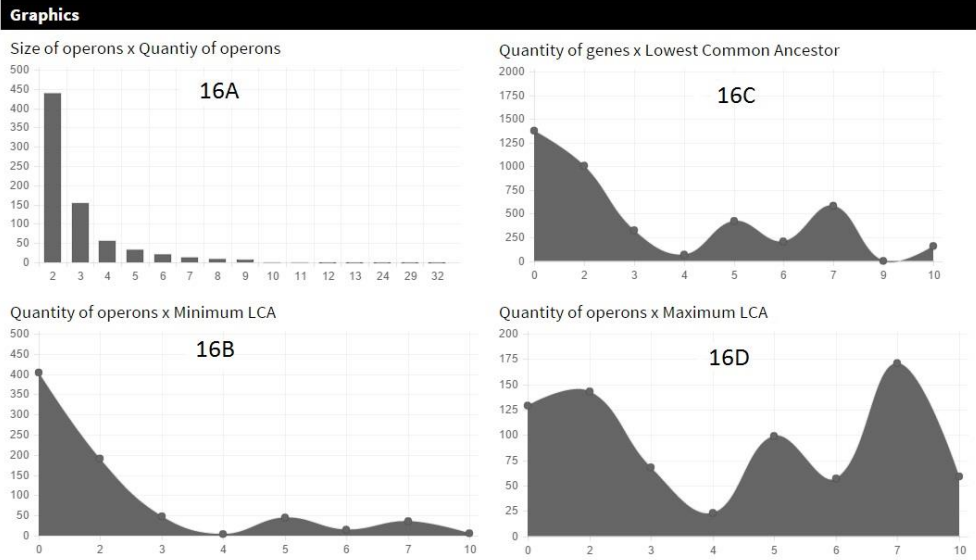


Figure 16: Graphics for organisms. The Graphics are automatically generated with MySQL, the queries were performed via PHP scripts and drawn using the Chart.js JavaScript library, publically available at www.chartjs.org. Graphic 16A shows the quantity of transcription units per size of transcription units. Part 16C shows the quantity of genes per speciation clade, 16B presents a graphic generated for the minimal LCA per TU, and 16D shows the maximum LCA per TU.

Using a random gene as example, Figure 17 presents the page for gene description with more specific information for gene concept. The page displays a complementation of the table of genes from the organism's page shown on Figure 15D. The data is complemented with GI number, position of gene inside the TU, strand, start and end base pair position relative to the first base pair position of the organism, gene sequence, and amino acid sequence with respective links for download.

BSU00310 (holB)

Main Information	
Gene:	BSU00310
Symbol:	holB
Protein:	NP_337912.1
GI:	1607099
Transcription unit:	23093309
TU Position:	3/8
Ortholog group:	460088
LCA (level):	cellular organisms (0)
Strand:	forward
Start:	40668
End:	41894
Organism:	Bacillus subtilis subsp. subtilis str. 168 (224300)

Sequence:

```

ATGGCAATATCTTGGAAAGAAATGAACGACCTTCACCCAGAGTGTGAAGCTTTTATTAATAGTATTGAAAGGACAG
ACTATCACGGCTTATTTTGGAGGGAAAGAGGACGGCAGCTGTGATCCGGCTCTTTGCAAAAGCTCTT
TTTCTGAGAGGTGGCCGGAGCCTTGTGAGAGTCCGAACTGTAAACGGATAGATCAGGAAACCTCGATCTC
CATCTTGTTCAGCTGTATGCTTATCATTAAGAAAGGCGCAATTCAGCGCTCCAGAGAGAGTTCCTAGACAGAGCT
TGAATGCGTAAAGAGTGTATATTTCCGACGCGGATTAATGAAAGCAATTCGGACAGAGCTCTCGAAATTT
TAGAAGAACCGAACAAAGACACATGGCCCTCTACTACTGACGACGCCCAAGGTTATGGATACCATCATTCAGGA
TGGCAAGCCTCTTTGAGCCTTTCAGCCGGAAGCAATGAGAGCAGATGTCGACAGAGAGGCTCCGCTCATAT
GGCAAGCCTTGGCCCAATGACTAATATATAGCAGAGCACTGAAATAGTGAATGATGATTTGAGAGTCTA
GAGCAAAAGTGAATAATTTGTGAAAGTCTACACAGCGGAAAGGACATCTTTTCTTTTCAAGATCAATGATG
GCTTTTTCAGAAAGAAAGCCGACAGAAATGGTGTGATATCTCTATGATATCTCGATGTGTGTGTGTGTGTGTG
AATAGAAATGAAATGAAATGATTTATACAGACTTATTCATCAATCAATAAACAGCATGCTGCTACATCAACACAA
GCCTTACAAATCAGTACTGTGTTTTAGAGCAAGAAAGCGCTTCAATGATGATGATGATGATGATGATGATG
CACCTGTGTTAATTTGTGAGAGGAGGATGA
  
```

Amino acid sequence:

```

MAISWKEHLEQPRVAKLLINSEVDRLSHAYLFEGKIVTGLDAALLAKSFPLEGGAEPCECRNKRHSQNPDL
HLVQDGLDKKVAIQALQEEFMTGLEHMLYSHADQITANANLLKLEPNDTHAVLITEPQLDITFSR
CGLPFGQFQHEZRLLECDQPHARLAKHNTVHVELESRVSEFESRHWLLELQKPRIGRHHFFDQVM
PPFKETHQEHGLMLLYRDLVLSIQIHEDLVYQDLFQSKQHALQSTQQQVTVQLAVLEAKRHLHSNINVQGLIE
HLLHLQEG
  
```

Download File

Figure 17: This screenshot presents the page for gene concept, with all information stored in the database for genes. The web page contains the basic information for each gene and links for transcription units and orthologue groups.

Following the concept of transcription unit, the page for this concept is shown in Figure 18, which is divided into two main parts: statistics of the TU and genes inside the transcription unit. The main statistics present the organism from which the TU comes from, the quantity of orthologue groups represented in that TU, and the maximum and minimum LCA present in the TU. The second part is similar to the gene information present on the gene table for organisms, shown on figure 15D, only with a small difference: the replacement of the TU web page link with the gene position on the TU.

Transcription unit: 28889309

Main statistics					
Organism:	Bacillus subtilis subsp. subtilis str. 168 (224308)				
Ortholog groups	8				
Maximum LCA (level)	Proteobacteria (3)				
Minimum LCA (level)	cellular organisms (0)				

Transcription unit has 8 genes					
ID:	Gene:	Symbol:	Ortholog group:	LCA:	Position:
36562	BSU00290	yaaQ	4545169	Firmicutes (3)	1/8
36563	BSU00300	yaaR	2910889	Bacteria (2)	2/8
36564	BSU00310	holB	460088	cellular organisms (0)	3/8
36565	BSU00320	yaaT	864911	cellular organisms (0)	4/8
36566	BSU00330	yabA	4461780	cellular organisms (0)	5/8
36567	BSU00340	yabB	845192	cellular organisms (0)	6/8
36568	BSU00350	yazA	5854776	Bacteria (2)	7/8
36569	BSU00360	yabC	456558	cellular organisms (0)	8/8

Figure 18: Displays the page of the transcription unit concept, divided into two parts. The first part shows the main information from TU and the second part is more related to the genes that form the operon, displaying links for the genes and orthologue groups with LCA for each gene and the position inside the TU.

The concept page structure of the orthologue group is divided into three main parts, being the first part related to basic information of the orthologue group, the quantity of genes clustered, and the LCA determined for that gene cluster.

The second part displays a graphic with the distribution of transcription unit's size of genes, which belongs to the orthologue gene cluster per quantity of transcription units with same size. The third part displays the description for each gene that belongs to the cluster. Figure (19) shows an example of the orthologue groups' page.

Ortholog group: 4545169

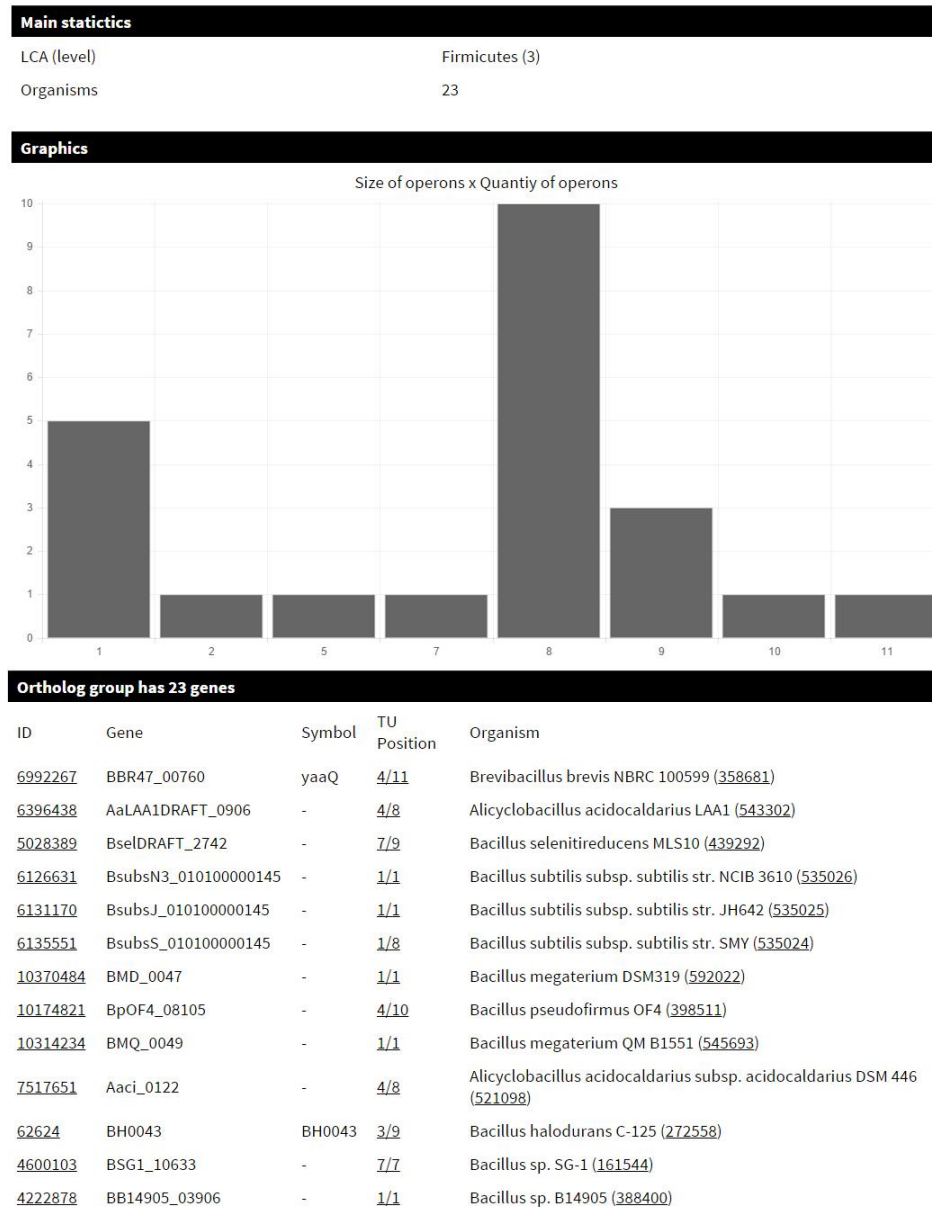


Figure 19: This figure presents the page for the orthologue group concept with the three main parts. The first one presents basic information of the orthologue group, the second part presents the graphic of size of transcription units per quantity of transcription units, and the last part presents the description of genes inserted in the gene cluster.

The next section of the TAXI's web interface presents queries of innovations existent on each organism; the innovations are divided in three main parts: innovations on transcription units, genes restricted to a specific clade, and transcription units restricted to a specific clade. Figure 20 presents a screenshot of the queries for taxonomy innovations.

Query

Query organism:

Chose one organism

Acaryochloris marina MBIC11017 (Taxonomy ID 329726) ▼

1. Show TU where

- First gene in TU is the most ancient
- First gene in TU is the most recent
- First gene in TU is more ancient than second gene
- First gene in TU is more recent than second gene

2. Show genes that are innovations restricted to

- species
- genus
- family
- order
- class
- phylum

3. Show TUs that are innovations restricted to

- species
- genus
- family
- order
- class
- phylum

Figure 20: Page for querying innovations on bacterial genome of a specific organism. There are three types of queries that can be performed. Query over innovations on transcription units, queries of innovations on genes restricted to speciation clades, and queries for transcription units restricted to a speciation clade.

A closer look on the types of queries reveals all options to customize them for innovations inside a specific genome. Below, the options for each type of query are described.

- First type: Presents transcription units in which;
 - First gene in TU is the most ancient – Query all transcription units of the organism in which the first gene is the most ancient of the TU;
 - First gene in TU is the most recent – Query all transcription units of the organism in which the first gene is the most recent of the TU;
 - First gene in TU is more ancient than the second gene – Query all transcription units of the organism in which the first gene is more ancient than the second gene, not mattering if both are the most recent or ancient of organism;
 - First gene in TU is more recent than the second gene – Query all transcription units of the organism in which the first gene is more recent than the second gene, not mattering if both are the most recent or ancient of the organism;
- Second type: Presents genes that are innovations restricted to;
 - The query for gene innovations can be restricted for: species, genus, family, order, class, and phylum;

- Third type: Presents TUs that are innovations restricted to;
 - The query for transcription unit innovations can be restricted for species, genus, family, order, class, and phylum.

TAXI's last section is the browse of organisms across all genome database. Figure 21 presents the browse options to perform the search for the desired organism, a small statistics about all genomes on the database, and a drop down menu to select the organism.

Browse

Main Statistics	
Organism:	351
Gene:	1490905
Transcription units:	779370
Ortholog groups:	223092

Choose one organism:

Acaryochloris marina MBIC11017 (Taxonomy ID 329726)

Figure 21: Presents the browse page, where a query can be performed for a desired organism and to access its complete statistics. The browse page is divided into two sections: a basic statistics of all genomes inserted in the database and a second section for choosing an organism for analysis.

2.4. Results and Discussion

The proposal of developing TAXI started in the idea of creating a database to compile the innovations on bacterial species across the cladogenesis process, in which every bacterium passed thought.

As previously discussed, the speciation process adds new features to the bacteria to adapt its mechanisms of growth and survival. These features could be added by a transferring process of a gene or of a complete set of genes from one bacterium to another.

TAXI try to present evidences using homologue gene clusters of how a set of genes were transported from one bacterium to another, evidencing process of the horizontal gene transfer through not phylogenetically related bacteria. However, since it is based on a very stringent orthologue clustering approach that comprises synteny amongst the criteria, in TAXI, the origin of a gene is understood as the origin in that scenario and context of synteny. Therefore, only horizontally-transferred TUs may cluster together.

The TAXI database was developed based on an already well-known biological database, the MicrobesOnLine database. Its genome annotations, operons prediction, and orthologue groups (MOGs) were used.

Starting with the publically accessible MicrobesOnLine database, only bacteria and archaea were selected for performing taxonomy innovations analyses, thus, supporting the evolutionary analysis by the community that studies these microorganisms. From a total of 3,707 genomes, the amount of 1,753 genomes was selected. Table 2 shows the total number of organisms, genes, transcription units, and orthologue groups used during the Taxonomy innovations analyzes.

Table 2: The table shows the number of organisms used in the TAXI database, with the total of genes transcription units and orthologue groups.

Concept	Organisms	Genes	Transcription units	Orthologue groups
Total	1.753	6.732.117	3.343.458	1.086.098

From the entire universe of bacteria used to create the TAXI database, Table 3 shows a total of 15 organisms with detailed quantities of genes, transcription units, and orthologue groups. Nine bacteria were used in the table: *Bacillus subtilis* subsp. *subtilis* str. 168 (Taxonomy ID: 224308), *Bacillus subtilis* subsp. *subtilis* str. NCIB 3610

(Taxonomy ID: 535026), *Bacillus subtilis* subsp. *subtilis* str. JH642 (Taxonomy ID: 535025), *Bacillus subtilis* subsp. *subtilis* str. SMY (Taxonomy ID: 535024), *Corynebacterium glutamicum* ATCC 13032 (Taxonomy ID: 196627), *Corynebacterium diphtheriae* NCTC 13129 (Taxonomy ID: 257309), *Escherichia coli* str. K-12 substr. MG1655 (Taxonomy ID: 511145), *Salmonella enterica* subsp. *enterica* serovar *Typhi* str. CT18 (Taxonomy ID: 220341), and *Shigella boydii* Sb227 (Taxonomy ID: 300268).

Six archaeal genomes were also used to be compared with the bacteria: *Aciduliprofundum boonei* T469 (Taxonomy ID: 439481), *Archaeoglobus veneficus* SNP6, DSM 11195 (Taxonomy ID: 693661), *Ferroglobus placidus* DSM 10642 (Taxonomy ID: 589924) *Halobacterium* sp. NRC-1 (Taxonomy ID: 64091), *Pyrococcus abyssi* GE5 (Taxonomy ID: 272844), and *Thermococcus sibiricus* MM 739 (Taxonomy ID: 604354).

Table 3: Table detailing quantities of genes, transcription units, and orthologue groups of a small set of organisms inserted in the TAXI database.

Organism	Taxonomy ID	Genes	Transcription Units	Orthologue Groups
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168	224308	4176	2612	4083
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. NCIB 3610	535026	4422	2480	4248
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. JH642	535025	4263	2644	4169
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. SMY	535024	4283	2442	4180
<i>Corynebacterium glutamicum</i> ATCC 13032	196627	3057	1993	2728
<i>Corynebacterium diphtheriae</i> NCTC 13129	257309	2272	1371	2001
<i>Escherichia coli</i> str. K-12 substr. MG1655	511145	4151	2439	4050
<i>Halobacterium</i> sp. NRC-1	64091	2075	1670	1989
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>Typhi</i> str. CT18	220341	4395	2446	4275
<i>Shigella boydii</i> Sb227	300268	4134	2713	3510
<i>Aciduliprofundum boonei</i> T469	439481	1544	1026	1267
<i>Archaeoglobus veneficus</i> SNP6, DSM 11195	693661	2193	1352	1785
<i>Ferroglobus placidus</i> DSM 10642	589924	2480	1382	2019
<i>Pyrococcus abyssi</i> GE5	272844	1780	1063	1710
<i>Thermococcus sibiricus</i> MM 739	604354	2035	1300	1787

Analyzing the data presented in Table 3, it is already interesting to note the difference of genome sizes between archaea and bacteria. It is easily noticed that, in general, the size of bacterial genomes are twice as bigger than archaeal genomes in terms of gene content.

This could be explained by the complexity of genomes and added features for supporting the bacterial survival.

To demonstrate the gene acquisition across speciation, TAXI's pipeline, using a web service publically available at *Biodados* laboratory, determined the LCA for each gene cluster available at the TAXI database.

Sending all organisms' taxonomy IDs whose genes are inside the same cluster, the web service responds the query with the common ancestor for that cluster, mapping all gene cluster of the TAXI database. Next, table 4 presents a general overview of the LCA for each organism on the database.

Table 4: Table presenting the quantities of genes divided per clade on the genome speciation of all genomes studied by the TAXI system.

Level	subspecies	species	genus	family	suborder	order	subclass	class	subphylum	Phylum	Kingdom	super kingdom	No rank	Total
	27.527	272.236	439.612	302.545	28.634	295.850	29.051	263.097	102.462	500.005	71.375	1.005.980	1.234.892	

The quantity of genes presented through clades shows two biases on the gene distribution. Ancient clades have a bigger quantity of genes and present a high quantity of shared genes amongst all studied genomes, and genes tend to be located on major clades than in subclades. Therefore, speciation is associated with the addition of fewer genes in more recent clades.

Executing the same analysis of Table 4, but showing the quantities for all fifteen previous mentioned organisms (Table 3), the results of the analyses were presented in the Table 5.

Table 5: Table presenting quantities of genes divided per clade over the genome speciation for all the previous mentioned organisms. The relation between Taxonomy ID and organisms which they represents is shown on Table 3.

Taxonomy ID	224308	535026	535025	535,024	196627	257309	511145	64091	220341	300268
No rank	1.376	1.397	1.365	1.371	932	769	1.473	899	1.457	1.377
super kingdom	1.007	1.016	993	1.000	961	502	997	76	937	872
Kingdom	-	-	-	-	-	-	-	-	-	-
Phylum	325	336	327	329	-	-	401	77	469	367
subphylum	-	-	-	-	-	-	-	-	-	-
class	68	73	69	69	64	47	234	-	249	187
subclass	-	-	-	-	61	70	-	-	-	-
order	423	421	413	414	449	281	-	-	-	-
suborder	-	-	-	-	82	49	-	-	-	-
family	207	206	204	204	-	-	916	722	820	1.184
genus	584	585	576	579	381	263	6	-	-	55
species	1	1	1	1	171	64	77	254	104	30
subspecies	158	301	298	294	-	-	-	-	339	-

Taxonomy ID	439481	693661	589924	272844	604354
No rank	957	1.205	1.352	1.015	1.134
super kingdom	123	160	193	169	158
Kingdom	-	-	-	-	-
Phylum	179	247	253	153	143
subphylum	-	-	-	-	-
class	-	-	-	-	-
subclass	-	-	-	-	-
order	-	-	-	-	-
suborder	-	-	-	-	-
family	-	157	278	-	357
genus	-	93	-	73	31
species	55	23	22	3	12
subspecies	-	-	-	-	-

By analyzing the data presented in the Table 5, the results show that the selected genomes follow the major patterns of gene acquisition, with a high quantity of genes being acquired by the organism on early clades than on newest ones.

It could also be observed on the results that related genomes tend to follow a pattern on gene acquisition with almost the same amounts of genes being added to the organism through speciation. This tendency is explained by the sharing of the same orthologue groups created in the newest clades and also shared by related species. It does not mean that the gene horizontal transfer process does not happen, but that it is a process that could only affect a lower number of genes at a time in which the process is being observed.

Moreover, Table 3 also shows that genomes with a bigger number of genes tend to spread the gene acquisition all over the speciation process. However, smaller genomes tend to concentrate the same process on ancient and recent clade levels, leaving a gap in medium clades, thus showing a clear difference between bacterial and archaeal genomes based on the size of the genomes.

This difference of gene acquisition distributions is not only observable in far

phylogenetic genomes, but it is also seen in bacterial speciation; all *Bacillus* genomes presented in table 3 follow a specific pattern of gene acquisition, differentiating the pattern of acquisition from distantly related bacterial genomes.

Comparing the four *Bacillus* genomes and *Salmonella* genome in Table 3 with other genomes examples, is denoted that those organisms have a higher level of speciation, since their process of cladogenesis ends in a more recent clade than other organisms.

This late gene acquisition is possibly an expression of a response to the environment in which the bacteria is inserted, presenting a higher grade of specificity of the genome that is acquiring genes or the entire new gene regulatory networks.

Genes that were acquired later are easily depicted in the database and they can be compared by other genomes using the orthologue groups to which the gene is inserted. Some examples of later acquired genes are presented in Table 6.

Table 6: Table showing a list of later-acquired genes, with general information about the gene.

Gene	Symbol	Protein	Transcription unit	Orthologue group	LCA	Organism taxonomy ID
<i>BSU01389</i>	<i>ybzG</i>	YP_003097669.1	28891879	6993271	<i>subspecies</i>	224308
<i>BSU01900</i>	<i>ybcM</i>	NP_388071.1	28889373	6318785	<i>subspecies</i>	224308
<i>b0135</i>	<i>yadC</i>	NP_414677.1	31583490	663471	<i>species</i>	511145
<i>b0280</i>	<i>yagN</i>	NP_414814.1	31583575	2585092	<i>species</i>	511145
<i>VNG0032H</i>	<i>VNG0032H</i>	NP_279192	28445313	3306493	<i>species</i>	64091
<i>VNG1838H</i>	<i>VNG1838H</i>	NP_280567	28446426	8012682	<i>species</i>	64091
<i>STY0010</i>	-	NP_454620	28860142	1230190	<i>subspecies</i>	220341
<i>STY0964</i>	<i>dmsC</i>	NP_455454	28860615	472373	<i>subspecies</i>	220341
<i>Aboo_0002</i>	-	YP_003482376.1	30836105	461305	<i>species</i>	439481
<i>Aboo_0220</i>	-	YP_003482594.1	30836265	4082729	<i>species</i>	439481

All genes in the example in Table 6 were collected from the highest level of each organism, reinforcing the concept that the speciation process varies between the studied genomes to which genes were added at any point of the speciation, not following a pre-prepared sequence of gene acquisition. It must be considered that the attribution of taxonomy classification might introduce a bias.

As presented in Table 6, each gene belongs to a transcription unit, which could be a

monocistronic transcription unit, to which the gene is transcribed alone, or a polycistronic transcription unit, to which the gene is transcribed with another gene or set of genes [75].

The analysis performed on the TAXI database with the orthologue groups also uses the information of operon predictions developed by the MicrobesOnLine database [75].

Some analyses of the transcription unit structure also used these predictions, studying in which period of the speciation the added genes on the organism were also added to a transcription unit.

The construction of a transcription unit could start with ancient clades and go through all speciation process of the organism, and finish on recent clades, following the entire process of speciation of the organism.

A transcription unit also might be built only on ancient clades, indicating that all genes of that transcription unit are shared with far phylogenetic genomes or were a basal TU with a basic function on the survival of any organism.

However, it is also possible to be a restriction to recent clades, where all genes of the transcription unit are only shared with close related genomes or these sets of genes might be exclusive to an organism, being a specific function for that organism.

With the use of the transcription units of genes on Table 6, some analyzes were performed on the structure of transcription units in which the genes are inserted. Since some units are monocistronic units, and to cover all possibilities of a transcription unit construction, there was the necessity of a complementation with more transcription units. All sets of transcription units are shown in Table 7.

Table 7: Table showing transcription units used for analyses of construction, studying in which speciation clade the first gene and last gene were added for the structure.

Organism Taxonomy ID	Transcription unit ID	TU size (in genes quantity)	Most Ancient LCA	Most Recent LCA
224308	28891879	1	<i>subspecies</i>	<i>subspecies</i>
224308	28889373	1	<i>subspecies</i>	<i>subspecies</i>
224308	28889374	6	<i>no rank</i>	<i>subspecies</i>
224308	28889404	2	<i>superkingdom</i>	<i>subspecies</i>
511145	31583490	7	<i>class</i>	<i>species</i>
511145	31583575	1	<i>species</i>	<i>species</i>
511145	31583552	4	<i>family</i>	<i>species</i>
511145	31583424	3	<i>no rank</i>	<i>no rank</i>
64091	28445313	1	<i>species</i>	<i>species</i>
64091	28446426	1	<i>species</i>	<i>species</i>
64091	28445448	3	<i>family</i>	<i>species</i>
64091	28445595	3	<i>no rank</i>	<i>species</i>
220341	28860142	2	<i>class</i>	<i>subspecies</i>
220341	28860615	4	<i>superkingdom</i>	<i>subspecies</i>
220341	28861065	2	<i>subspecies</i>	<i>subspecies</i>
220341	28861084	2	<i>subspecies</i>	<i>subspecies</i>
439481	30836105	1	<i>species</i>	<i>species</i>
439481	30836265	1	<i>species</i>	<i>species</i>
439481	30836107	5	<i>phylum</i>	<i>species</i>
439481	30836326	2	<i>phylum</i>	<i>species</i>

Table 7 presents twenty transcription units, ten from the Table 6 complemented with more ten units with more than one gene in TU. Besides the transcription unit ID, there are also presented the size of the transcription unit, and the most ancient and most recent LCA for each transcription unit. This depicts the moment of origin and the accomplishments of the TU composition. Some TUs might encompass a period encompassing from phylum to species, while others may come from between family and species.

As mentioned early, all cases of transcription unit construction are presented in Table 7: transcription units with all genes of early clades, transcription units following the genome speciation, lasting in all clades, and transcription units formed by genes acquired in clades that are more recent. Figure 22 presents examples of transcription unit structures from Table 7.



Figure 22: Figure graphically presenting all transcription units of table 7 with a key linking the colors on the graphic with the clades of speciation divided by organism.

Starting with monocistronic transcription units, this type of TU is present in any clade, since they are observable introductions of unique genes to the organism genome at any point of the speciation process. Examples of this type of TU are: 28891879, 28889373, 31583575, 28445313, 28446426, 30836105, and 30836265.

An example of transcription unit completely formed by genes with ancient LCA is the TU ID number 31583424 of the *Escherichia coli* str. K-12 substr. MG1655 bacteria; it is formed by three genes that are shared with other cellular organisms, proving that the genes of this transcription unit could be found in far phylogenetic organisms.

In the examples on Table 7 and Figure 22 there are also transcription units, whose building process encompassed the entire speciation of the organism with genes being added on early clades and finishing the construction on newer clades. The major example of this kind is shown on TU ID 28889374, with genes shared with other cellular

organisms, genes that, as mentioned, could be found in far phylogenetic organisms, with genes added on newer clades such as order and subspecies, showing a modification on the transcription unit that is exclusive for the bacteria *Bacillus subtilis* subsp. *subtilis* str. 168.

The last type of transcription unit structure is presented on the examples 31583490 and 31583575 of *Escherichia coli* str. K-12 substr. MG1655, 28445448 of archaea *Halobacterium* sp. NRC-1, 28860142, 28860615, 28861065, and 28861084 of *Salmonella enterica* subsp. *enterica* serovar *Typhi* str. CT18, and from the archaeal genome of *Aciduliprofundum boonei* T469, the example 30836107. All these samples show a certain degree of exclusivity with the organization of transcription units already starting on newer clades, being shared therefore only with other near phylogenetic bacteria.

Transcription units that are totally formed by genes with ancient LCA could be explained by a set of genes that develop an important role on the survival of organisms, or could be explained by the process of horizontal gene transfer, where the complete transcription unit is transferred between two or more organisms that share the environment or function.

Mixed transcription units with genes that have an ancient LCA and genes with recent LCA create the hypothesis that an old transcription unit gained new functions during the speciation process, in which some genes were clumped to the TU, in which the new structure of the transcription unit could be exclusive for these bacteria. Moreover, it shows that the organism is in a constant evolution.

Totally-recent transcription units present a response of the organism to an internal or external modification, which can be, for example, a response to a new environment where the organism was inserted, adjusting the products of the cell for its needs.

The modification of transcription units could generate diversity between TUs sizes across organisms; a transcription unit could insert or delete a gene in another organism due to biological pressures.

With the transferred transcription units by horizontal gene transfer, other transcription units could be generated by biological pressure, since the genes are transcribed together satisfying cell needs. This biological pressure generates new transcription units and this could generate diversity of TU average size through

organisms. Figure 20 presents a graphic comparing the size of transcription units per quantity of genes present in the TAXI database.



Figure 23: Graphic comparing the size of transcription units per quantity of calculated transcription units, considering all organisms inserted in TAXI database.

The range of transcription units inside the TAXI database varies between 2 and 69 genes per TU, the most common size is two genes per TU, more than the double of three genes per TU. The quantity of transcription units with size above ten is quite insignificant, while there is only few TU per size.

The results of the quantity analyses of transcription units per TU size for specifically analyzed organisms are presented on Figure 24, which compare the structure of transcription units through different taxa chosen just for comparison.

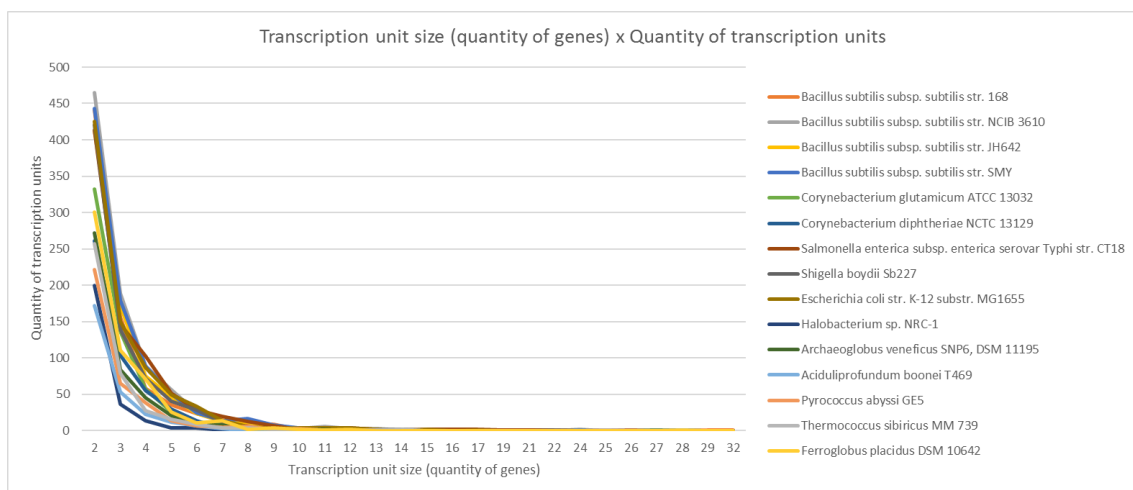


Figure 24: Comparison of transcription unit size of fifteen analyzed organisms covering a vast quantity of taxons, presenting the similarity and differentiation of related and unrelated organisms. Each curve presents an analysis for one specific organism, following the same pattern observed on Figure 23 of the transcription unit size distribution.

Figure 24 presents the curves that represent the correlation of transcription unit sizes with the quantity of transcription unit with the same size. Each curve shows a similarity with the major pattern found in Figure 23, with a fast decrease in quantity of transcription units in smaller sizes and a long tail of big TUs, with a big quantity of genes per transcription unit.

Dividing analyses in groups for a best understanding of correlation of TU size per quantity shows the similarity between phylogenic related organisms and the difference between not related organisms; Figures 25, 26, and 27 show respectively the distribution between transcription unit sizes in groups.

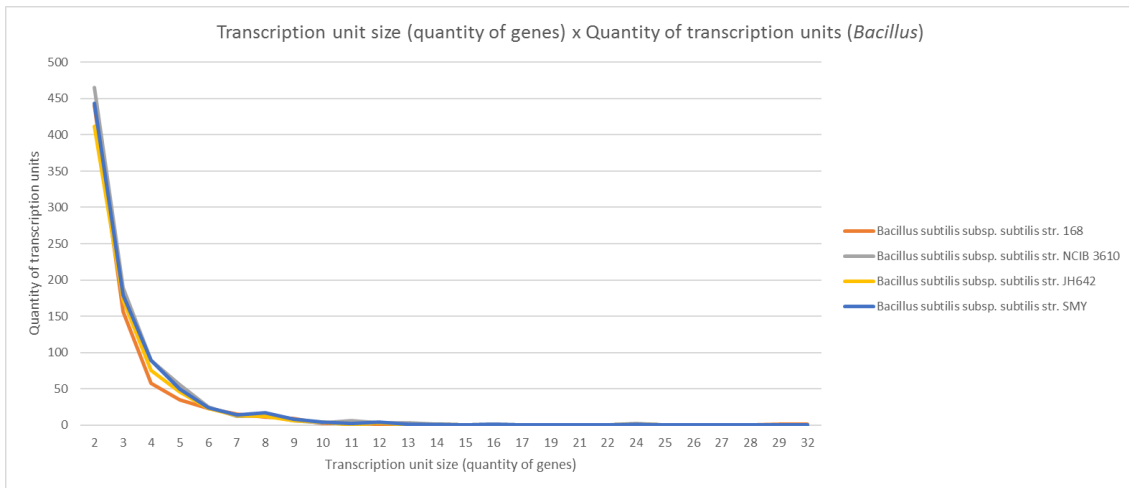


Figure 25: Figure presenting the correlation between transcription unit sizes per quantity of transcription units with the same size. It is the same analysis presented on graphic 24, but only for *Bacillus*.

Figure 25 presents the analyses for *Bacillus* bacterial genomes. All data from *Bacillus* follows the major pattern, but as previously mentioned, more related organisms tend to preserve the pattern of the transcription unit's distribution size, presenting a little difference on the curves. Here, the bacterium *Bacillus subtilis* subsp. *subtilis* str. NCIB 3610 (Taxonomy ID: 535026), *Bacillus subtilis* subsp. *subtilis* str. JH642 (Taxonomy ID: 535025), and *Bacillus subtilis* subsp. *subtilis* str. SMY (Taxonomy ID: 535024), more related bacteria, are grouped together, besides the bacteria *Bacillus subtilis* subsp. *subtilis* str. 168 (Taxonomy ID:224308), which presents a different curve of transcription unit sizes.

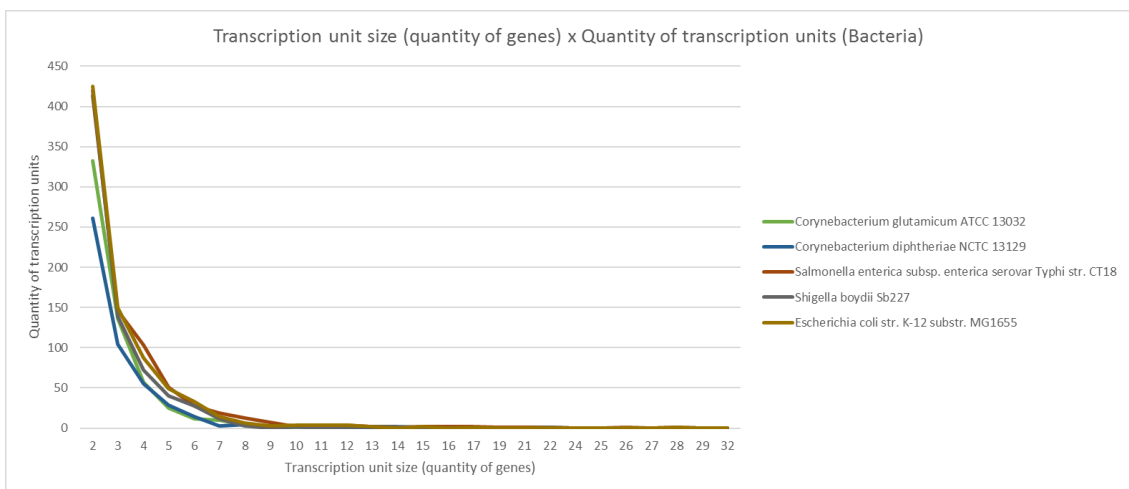


Figure 26: Analyses performed for other group with five bacteria, showing the curves for transcription unit sizes per quantity of TU. This figure presents a similarity between the curves of bacterial genomes.

All bacteria presented in this other group (Figure 26) shows almost the same pattern for operon sizes, with the biggest similarity shown by *Escherichia coli* str. K-12 substr. MG1655 (Taxonomy ID: 511145), *Salmonella enterica* subsp. *enterica* serovar *Typhi* str. CT18 (Taxonomy ID: 220341), and *Shigella boydii* Sb227 (Taxonomy ID: 300268), whose curves generated on the graphic show almost a complete overlap.

The last two bacteria *Corynebacterium glutamicum* ATCC 13032 (Taxonomy ID: 196627) and *Corynebacterium diphtheriae* NCTC 13129 (Taxonomy ID: 257309) also present an overlap in their curves because they are bacteria from the same genus and share a considerable number of transcription units.

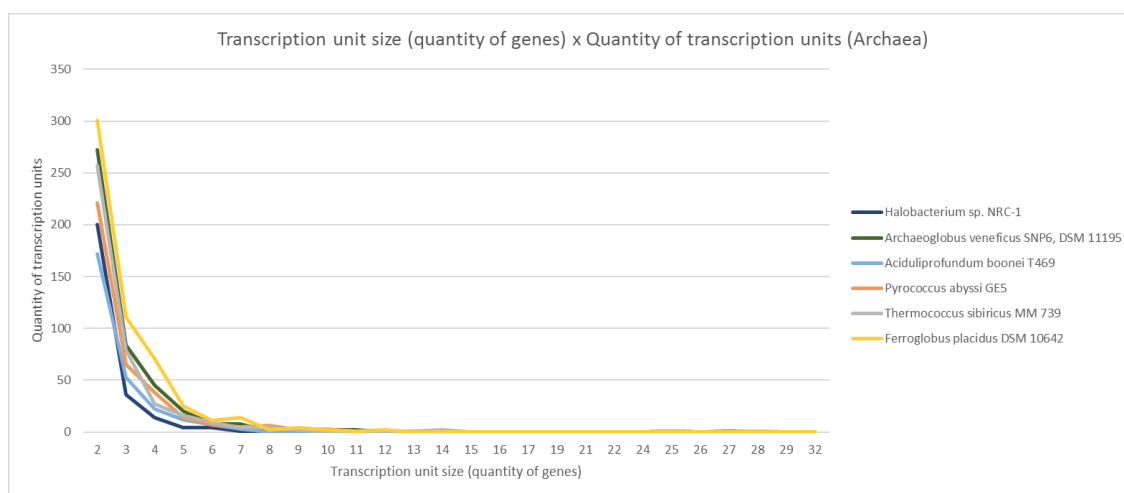


Figure 27: Figure presenting the curves of transcription unit distribution per transcription unit size for *archaea*. Five out of six organisms presented on the graphic show almost the same curve structure, except for the *archaea* *Ferroglobus placidus* DSM 10642.

The group of *archaea* (Figure 27) presents almost the same structure of all organisms present on the group, except for the *archaea* *Ferroglobus placidus* DSM 10642 (Taxonomy ID: 589924), which presents a different curve of transcription units from the pattern shown by other *archaea*. This difference might be explained by the size of the genome of the organism.

All other *archaea* presented on the graphic show the same pattern of distribution of transcription unit size because they present comparable gene quantities amongst the organisms.

The difference of curves is represented in Figure 28, where a set of six organisms are compared: four bacteria *Bacillus subtilis* subsp. *subtilis* str. NCIB 3610 (Taxonomy ID:

535026), *Bacillus subtilis* subsp. *subtilis* str. 168 (Taxonomy ID: 224308), *Corynebacterium glutamicum* ATCC 13032 (Taxonomy ID: 196627), *Escherichia coli* str. K-12 substr. MG1655 (Taxonomy ID: 511145), and two *archaea* *Ferroglobus placidus* DSM 10642 (Taxonomy ID: 589924) and *Halobacterium* sp. NRC-1 (Taxonomy ID: 64091).

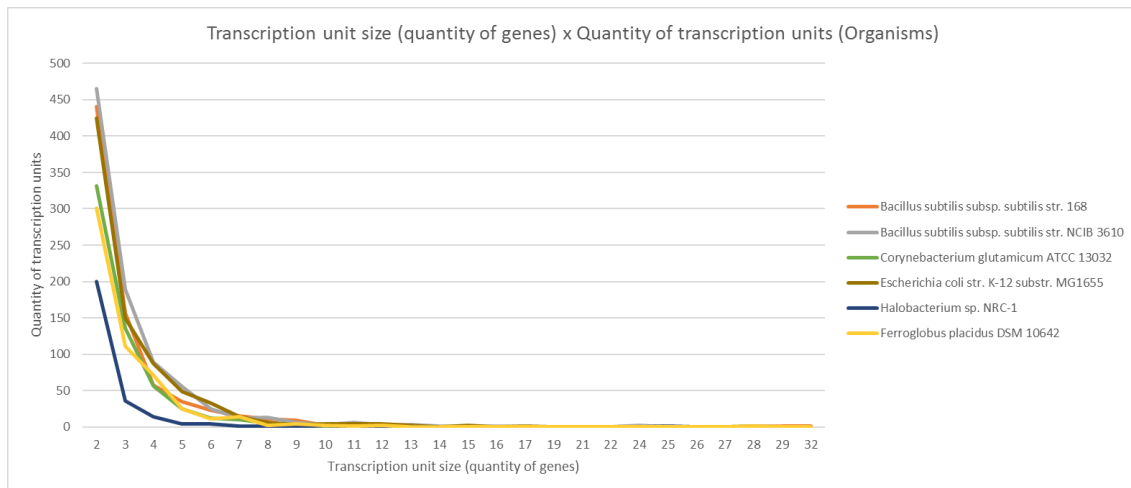


Figure 28: Comparison between *Archaea* and *Bacteria*, showing a difference of the transcription units' structure patterns of *archaea* *Halobacterium* sp NRC-1, which show a different pattern that is shared with other archaea, except for the archaea *Ferroglobus placidus* DSM 10642, which shows a curve similar to the *bacteria*.

2.4.1. Gene classification through bacterial speciation

As previously discussed, genes are shared amongst all clades, between phylogenetically organisms, related or not, presenting genes that are more related with the basal functions of the organisms or with specific features added on the genome as a response for internal or external triggers.

Starting with the examples used in the previous section, which were shown in this section's Table 3, the evolution of the organisms will be further discussed, presenting all clades of each organism and the quantities of genes added to it respectively, presenting all steps of the organisms through the speciation process.

Executing the same analyses as previously, but including all clades and subclades participant of the speciation process, not only the major clades as in the previous section, with the goal of going through the entire phylogeny of the organism, the basal and exclusive genes for the genome selection from the studied organisms' universe will be

presented.

Using the bacterial genomes of all four *Bacillus* presented as examples, which are listed 29, the results of the analyses of mentioned organisms are presented with all clades of this bacteria speciation.

Quantity of genes added per clade

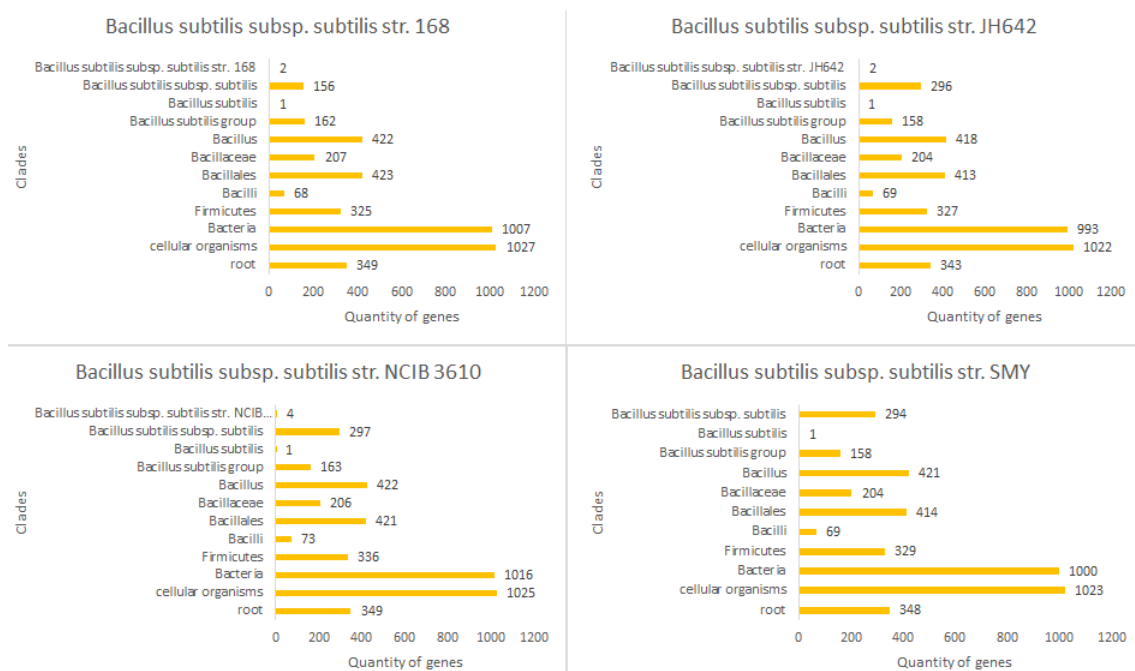


Figure 29: Comparison between all *Bacillus* bacterial genomes, amongst every clade of the bacterial speciation. The quantity of genes shared in each clade is compared.

Figure 29 presents four *Bacillus* bacterial genomes; the same examples used previously. The graphic shows the quantity of the genes shared with other organisms per speciation clade or the genes exclusive to a species, subspecies, and strain.

All four *Bacillus* presented a similar result, since they share almost the same evolution process with little divergences between them. The remarkable modification is the absence of the “strain exclusive genes” for the *Bacillus subtilis* subsp. *subtilis* str. SMY.

The peaks of gene acquisition for all *Bacillus* presented in Figure 29 are, on ancient clades, showing that the speciation of these bacteria have a high level of shared genes with far phylogenetically organisms. It also shows that the majority of survival mechanisms is present in other distant organisms.

Next, Figure 30 compares a set of bacteria chosen to cover two important bacterial genus; for each bacteria, its lineage was followed, so the different clades of non-related bacteria present the differentiation of this bacterial speciation. Only clades with gene acquisitions are shown.

Quantity of genes added per clade

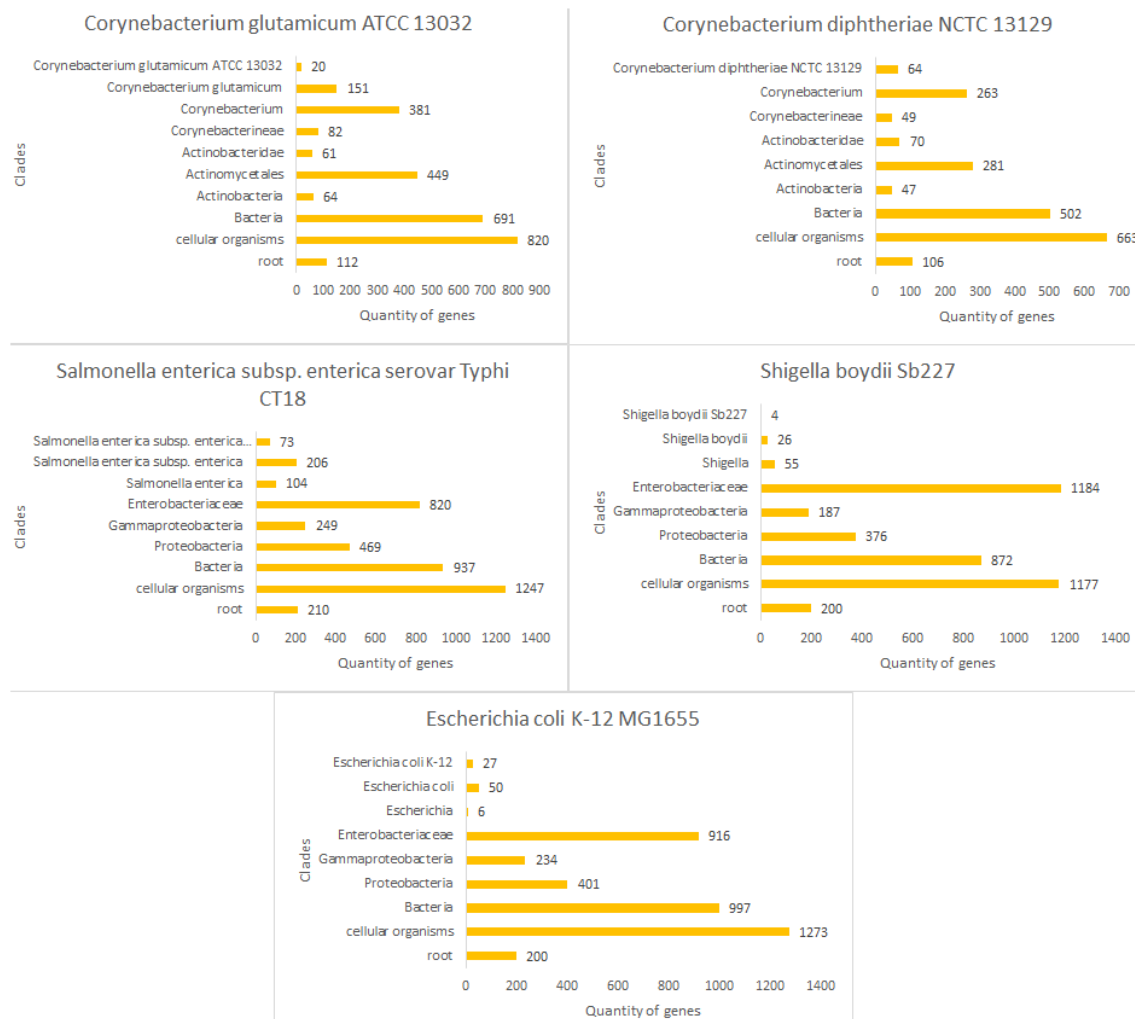


Figure 30: The figure shows the speciation of five bacteria among the clades, showing the gene acquisition for each clade. It presents the similarities and differences between different species, varying the length of the speciation process and the number of genes acquired.

On Figure 30, it is possible to see the similarities and differences amongst the speciation processes through which each studied bacteria passed through. It is easy to discern that there are two major epochs of gene acquisition for every bacterium, but all share the large gene gain in the cellular organism clade. These genes are classified in this clade because they are shared, either as archaea or eukaryotes.

The second peak of gene acquisition varies with the studied bacteria. For both *Corynebacterium* bacteria, the genus clade is the second biggest peak of gene acquisition, representing a big number of genes being acquired by all bacteria that belong to that genus.

Besides the *Corynebacterium*, the other bacteria share a high amount of genes on the family clade, presenting an earlier acquisition of features when the bacteria becomes an *Enterobacteriaceae*; this could be explained by the period, during the speciation, when the bacterium required a response for new environments.

For other *bacteria*, it is completely plausible that the peaks of gene acquisition occur in different steps of the speciation. In all examples presented here, they can be separated in groups with different clades with peaks of gene acquisition.

As for *bacteria*, the *archaea* organisms also present some patterns of gene acquisition. In other words, they also have specific peaks of genes acquisition, but since the *archaea* have a smaller number of genes, some organisms present only one peak of gene acquisition, proving that important survival mechanisms are shared with distant phylogenetic organisms. Figure 31 shows a comparison amongst six *archaea* with same analyses of gene acquisition presented for previous *bacteria*.

Quantity of genes added per clade

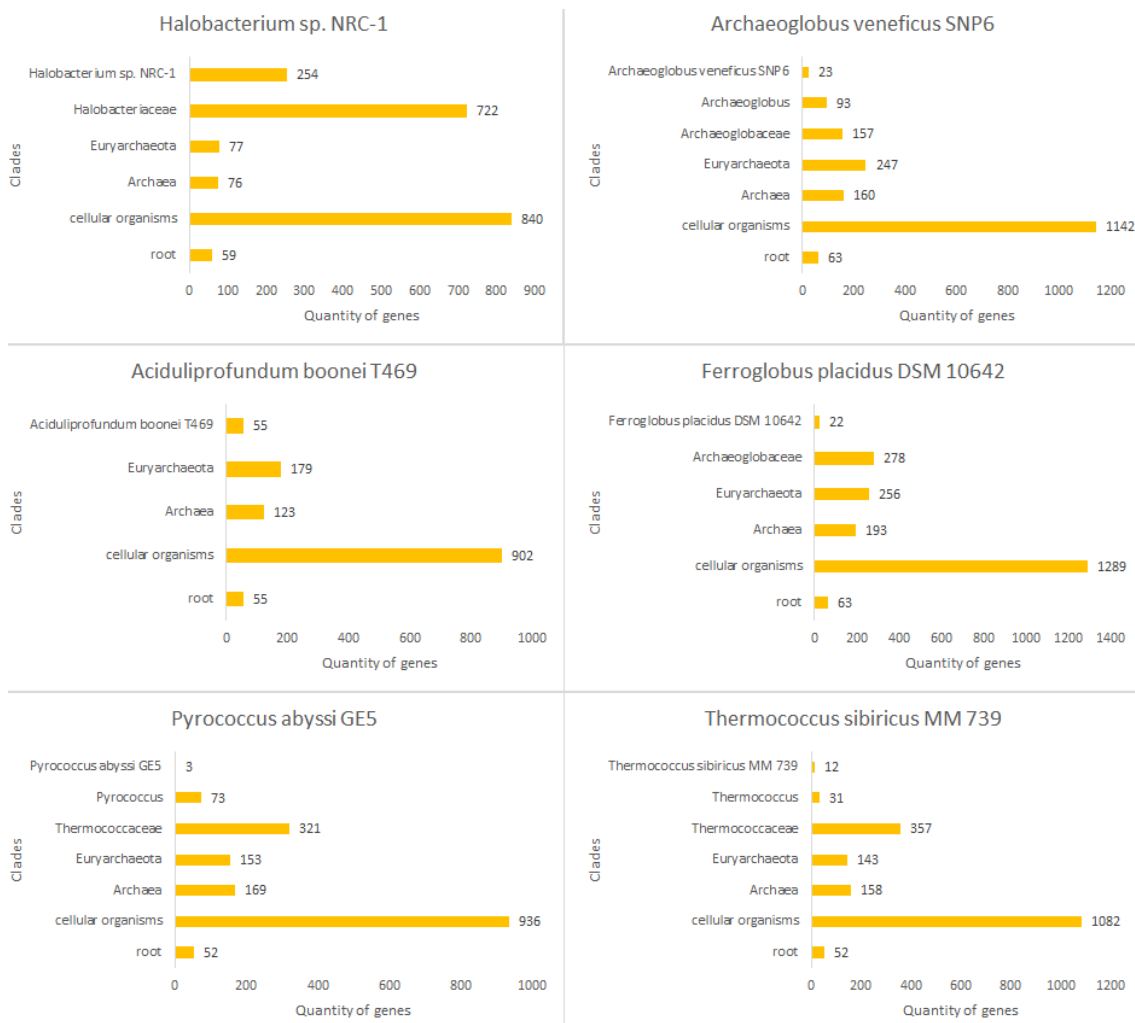


Figure 31: Graphics comparing the speciation of six Archaea, showing different peaks of gene acquisition in comparison with the patterns presented by previous discussed bacteria.

Comparing with bacteria, almost all studied archaea present a different pattern of gene acquisition. The exception is seen on the archaeal genome of *Halobacterium* sp NRC-1 (Taxonomy ID: 64091). This archaeon presents two peaks of gene acquisition, a pattern seen on bacterial genomes.

The other studied archaea present only one peak of genes acquisition. This peak is located on a previous clade, as the first peak presented on bacterial genomes. This peak, for all organisms, presents a set of genes that are basal for all studied organisms, composing a set of genes that supports the survival of cellular organisms.

Also, as observed in bacteria, archaeal genomes present pattern similarities of gene acquisition for related organisms; the genomes of *Pyrococcus abyssii* GE5 (Taxonomy id:

272844) and *Thermococcus sibiricus* MM 739 (Taxonomy ID: 604354) present a high similarity of gene acquisition with a small peak on the family clade with almost the same quantity of genes acquired per clade.

The next section will discuss how these genes acquisitions behave on construction of transcription units, how TUs were formed, and when, during the speciation process of the example organisms, the TU composition started and finished.

2.4.2. Construction of transcription units through bacterial speciation

As previously discussed, gene acquisition occurs during the speciation process of the organisms, and transcription units are a mechanism used by the cell to produce a set of proteins at once, responding to only one stimulus.

The transcription units are a set of genes transcribed as one that share the initiator and the terminator. As previously mentioned, the transcription units are conformed by a set of related genes that can be added to the TU at any moment during speciation, as far as the construction can start in early clades and stop on later ones, as shown, for example, in Figure 32.

This acquisition of genes by existing transcription units could be forced by biological pressure or mutations occurred in the genome. Paralog genes or mutated genes, which can be specific to one species and participate later on a process of horizontal gene transfer, are possibly added to operons of non-related phylogenetic organisms.

Dividing the analysis among the example organisms, Figure 32 shows the structure of transcription units for the group of *Bacillus* organisms previously shown on Table 3. Besides the tendency of starting the operon in the ancient clade of the cellular organism, there are some operons that are being generated more recently. Moreover, the tendency for finishing is higher in clades that are from a more recent genus for these bacteria.

Acquisition of genes through cladogenesis

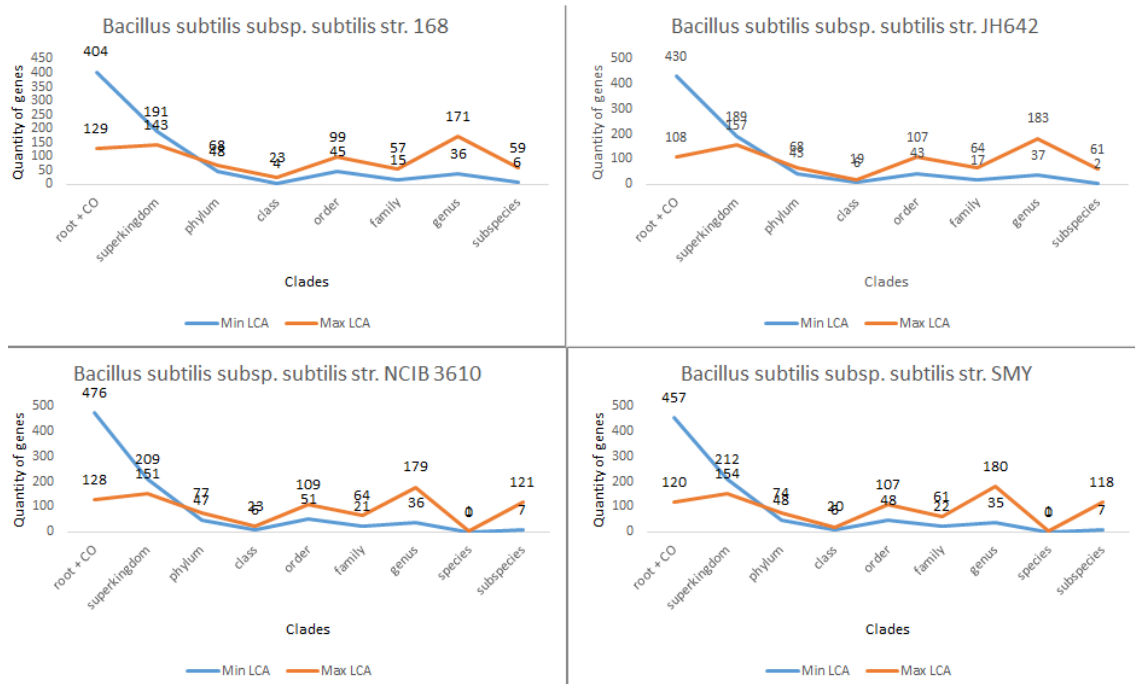


Figure 32: Transcription unit structure for *Bacillus* organisms shown in Table 3. The lines for minimum LCA represent the quantity of transcription unit constructions started per clade and the maximum LCA lines show the quantity of transcription unit constructions ended per clade, i.e., the ancestry for the first and the last genes added to the operon, respectively.

The pattern of transcription unit structure shown by the graphics in Figure 32, made available for all organisms on the TAXI database, follows the pattern presented in analyses for all four *Bacillus*. Again, for this small group, the start of the transcription unit composition is especially focused on early clades and the end of TU construction is spread all over taxonomy clades with small peaks on recent clades.

Comparing the inside of these four example organisms, there is the addition of only one gene in all four organisms on the species clade, but for TU construction, only in the organisms *Bacillus subtilis* subsp. *subtilis* str. NCIB 3610 and *Bacillus subtilis* subsp. *subtilis* str. SMY. This gene is used for finishing the TU structure when the gene is inserted.

The results of the same analyses for other groups of bacteria, with five example organisms covering the *Proteobacteria* and *Actinobacteria* families, are presented in Figure 33.

Acquisition of genes through cladogenesis

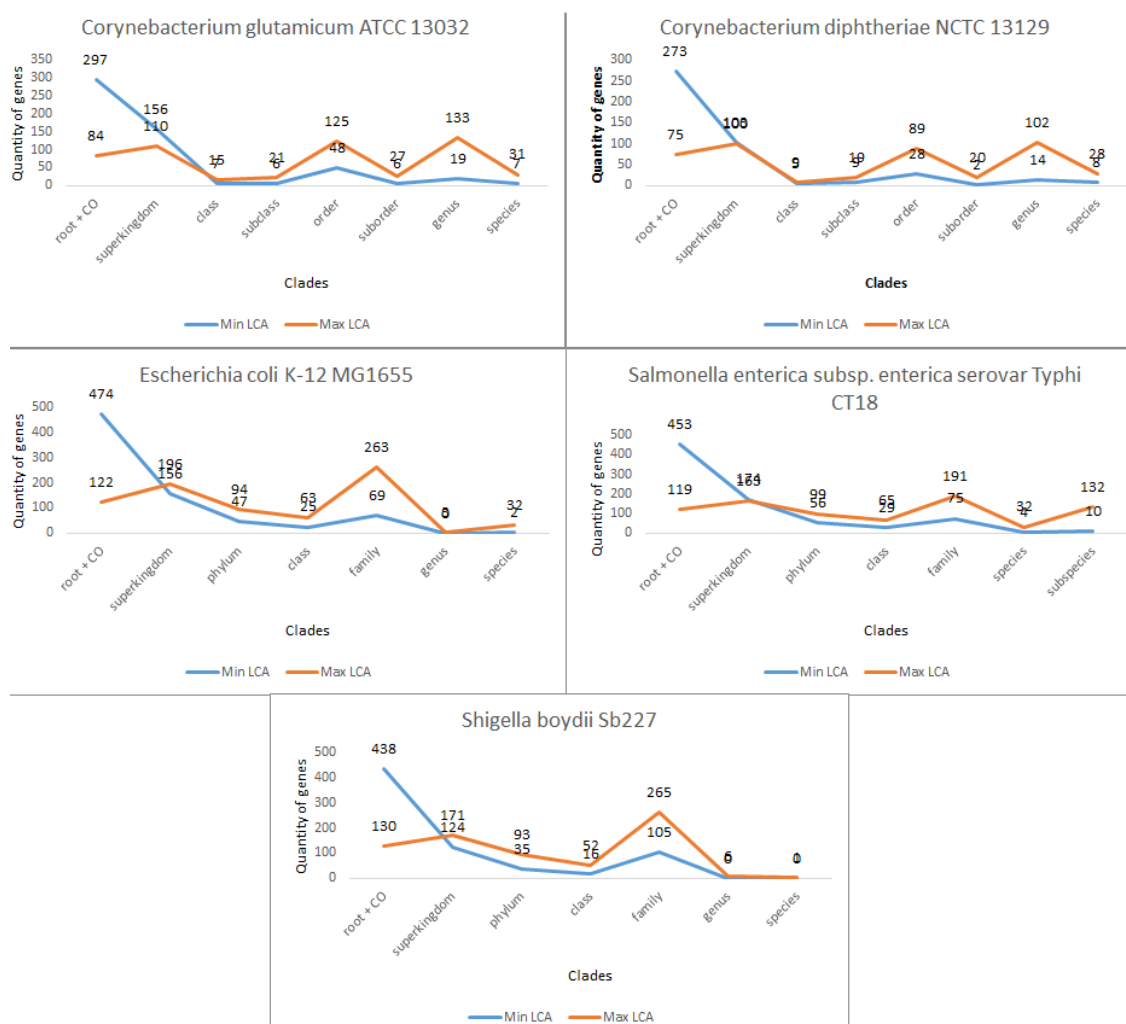


Figure 33: Analyses of transcription unit construction for five bacterial genomes divided into *Proteobacteria* and *Actinobacteria* families, comparing the process through the clades of the organisms.

Comparing the presented in Figure 33 with previous results for *Bacillus*, it shows the same pattern for starting the construction of transcription units, but there are differences in the line that presents the results for ending the process. Therefore, there are recent additions to the existing operons and they vary between different bacteria. This reinforces the necessity of a comprehensive database for many bacteria, which are represented in TAXI.

As for the finishing of operon construction, a comparison of the two previous results show that they could fit in two groups: *Bacillus subtilis* subsp. *subtilis* str. 168, *Bacillus subtilis* subsp. *subtilis* str. JH642, *Corynebacterium glutamicum* ATCC 13032, and *Corynebacterium diphtheriae* NCTC 13129 can be suitable in one group, presenting a bigger similarity than with other organisms. The pattern shows two separated peaks on

recent clades, representing two moments of adaptation.

All other bacteria create another group: *Bacillus subtilis* subsp. *subtilis* str. NCIB 3610, *Bacillus subtilis* subsp. *subtilis* str. SMY, *Escherichia coli* K-12 MG1655, *Salmonella enterica* subsp. *enterica* serovar *Typhi* CT18, and *Shigella boydii* Sb227. They have a different line of results for finishing the construction of transcription units; this groups present only one higher peak in one recent clade, but not in the same clade for all organisms; for *Bacillus* genus, the peak is exactly on genus clade, but for *Enterobacteriaceae*, for all three bacteria, the peak is on the family clade.

Moreover, for the last organisms, the same analyses were made: six *archaea* were analyzed referring to the construction of transcription units, as it was previously performed for bacteria. Figure 34 shows the results of this analysis.

Acquisition of genes through cladogenesis

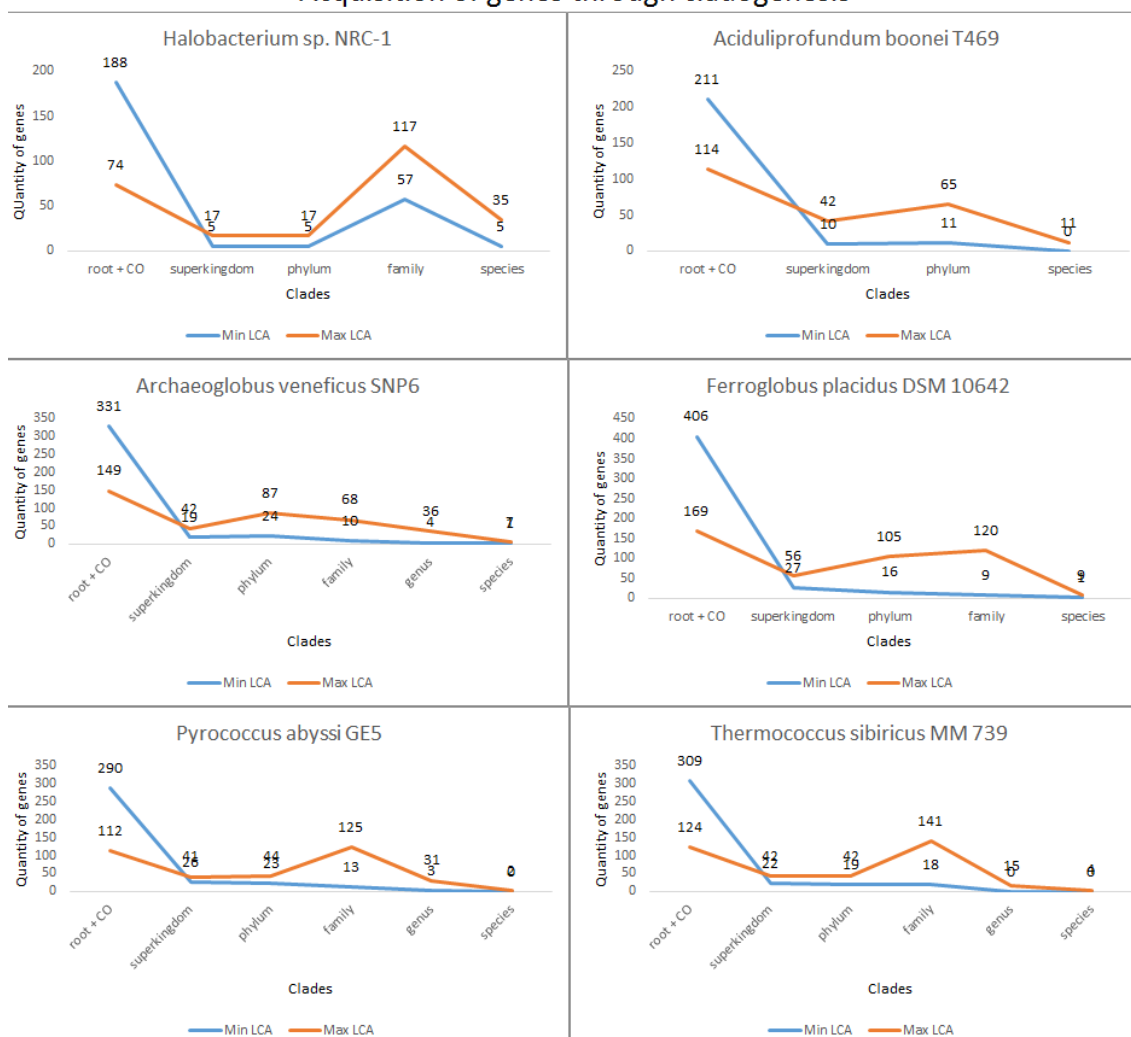


Figure 34: Figure displaying the analyses for *archaea*. The analyses were performed using the same process as for bacteria, as discussed previously. With less genes, the speciation process was influenced and it presented a different pattern for the structure of transcription units.

With fewer genes in the genomes, the archaeal genomes present a different pattern for structure of transcription units; the modification is seen in both lines.

As for bacteria, the last clade of gene addition of transcription units is also spread all over the speciation clades of the organisms, although with the peak difference shown in some examples.

Halobacterium sp. NRC-1 shows a totally different curve for starting the TU structure process. It presents two peaks: on the first clade, root and cellular organisms, and on family, showing a high TU acquisition on recent clades, which could be explained by the acquisition of new features by the organism that survives on a new environment.

All other organisms express, in the starting line, almost the same pattern observed in bacteria: a high peak on the first clade and a long tail for later clades with a small amount of cases of starting new transcription units, although the origin of new operons is rarer.

For bacteria, the curve for finishing the TU construction process shows different patterns. For *Halobacterium* sp. NRC-1, *Pyrococcus abyssi* GE5, and *Thermococcus sibiricus* MM 739, it is easy to notice a high peak of recent LCA of TUs, which means that there was a great gene acquisition on old transcription units, which could be explained by the biological pressure on clustering-related genes or mutated genes.

Aciduliprofundum boonei T469, *Archaeoglobus veneficus* SNP6, and *Ferroglobus placidus* DSM 10642 show a dissimilar pattern for finishing the construction process, a more spread one on the clades, but not as the one observed for bacteria. With a high peak on early clades and a small peak on the later ones, the ending of structure process apparently can be more comparable to the starting lines of other organisms. This could be explained by a small amount of genes in the organisms and the use of highly-shared features with far phylogenetic organisms.

2.5. Conclusion

The use of publically available databases, MicrobesOnLine, RegulonDB, KO, UEKO, for example. Created the basement for the development of this project, rising a new database for studying the cladogenesis process of bacteria and archaea.

All process of creating this new system, creating the database based on information collected from MicrobesOnline, such as, operons prediction and orthologue groups, followed by the calculation of Lowest common ancestor for all genes, and consequently the most ancient and most recent gene of each transcription unit. Created the basement or our study. With information collected and grouped into a database.

After the creation of the database, it was extremely important to create the front-end of the database to present the results, and allow the user to perform analyses and comparisons of the data of bacterial and archaeal organisms.

The analysis of the stored data addresses the differences between organisms. Gene acquisition in phylogenetically-related organisms tends to follow a pattern of quantities of genes added in each clade. On the other hand, far phylogenetic organisms may even possess shared genes, but they present a very different pattern of gene acquisition. Although some comparisons have been presented here, we realize that researchers interested in the speciation process of a given family, genus, etc., may make a specific use of TAXI to support their studies.

A complete set of genes could be shared between an organism phylogeny, a transcription unit. The transcription unit may have the same structure between two related organisms, but as shown in the results, even related organisms can evolve differently, responding to different external biological pressures. This difference may change the structure of transcription units, adding or deleting new components to the units.

The comparisons made between all organisms studied may facilitate the understanding of bacterial and archaeal evolution, simplifying the study of differences generated by different environments.

3. CORYNEREGNET 7.0

3.1. Aims

3.1.1. Main

Starting with one existing system, a work to update and adapt new features to CoryneRegNet began, presenting a new and easier way to add new organisms or update information already stored in the database. With this work, a new front-end was implemented also; it is more user-friendly with new features, making the access to information handier to researchers.

3.1.2. Specific

To develop a new concept-based data structure that might faster support all data stored in the database, supporting the growth of the database with the addition of new organisms or regulations.

To create one entirely new back-end developed in Java, using widespread libraries for a faster and more reliable database creation, updating the transfer of regulation pipelines, and covering old and new concepts of transferred regulations.

To update the data source with new validated data convenient to chip-seq analysis for transcription units, adding also the concept of sub transcription units and new validated data for binding sites of transcription factors and sigma factors.

To modernize the front-end, adapting new technologies and improving the researching experience of the available data in CoryneRegNet, using new approaches for the present data, eliminating the distinction between validated and predicted data, using only one database.

3.2. Related work

3.2.1. PRODORIC

PRODORIC (<http://www.prodoric.de/index.php?index=1>) is a database developed and published by the Bioinformatics Competence Center of Braunschweig that describes a great number of controlled vocabulary of annotated information on the regulation of gene expression in prokaryotes [27].

This database was created in 2003 as a universal data source covering gene regulation in prokaryotes, focusing in pathogeny. The data sources are populated via constant manual curation of scientific literature, whose main part contains a unique collection of binding sites of transcription factors and their interacting transcription factors. The transcription start sites are also included in the database with related sigma factors. In its database, PRODORIC includes a total of 2,921 binding sites of transcription factors and 197 position weigh matrices [27].

As previously discussed, PRODORIC's tools also present a web interface, as shown in Figure 35, which allows the user to perform analyses of their database. Using an associated tool, Virtual Footprint, the user can perform a query in an entire bacterial genome, searching for predicted TF-DNA interactions [76]. Although PRODORIC excludes all computationally-predicted data from the database, prediction analyses are possible by the front-end with the use of the constantly-updated database as a resource for predictions [27].

Also, on the web interface shown in Figure 35, a graph visualization tool, ProdoNet, is available. It explores PRODORIC's contents; this tool is capable of creating a graphic view of gene regulatory networks in multiple levels, like regulatory circuits and several network motifs [77].

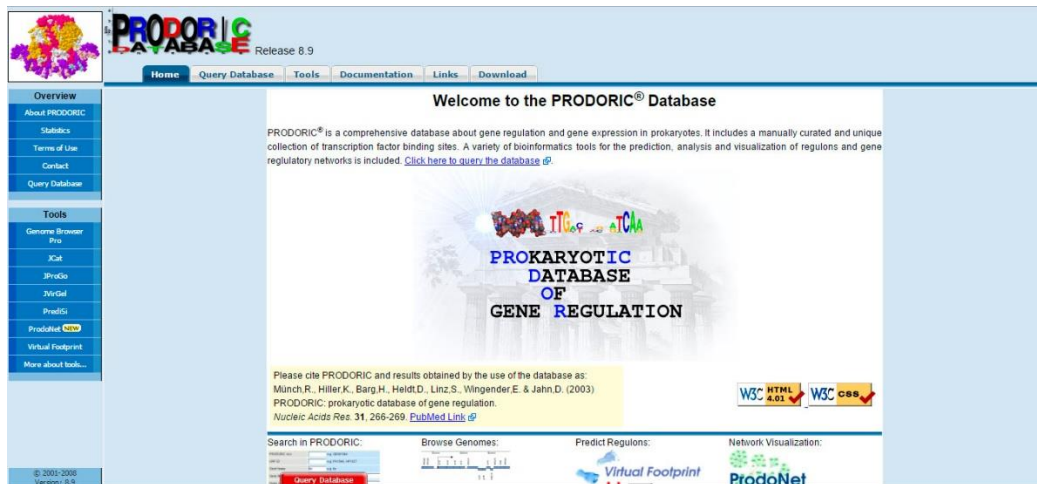


Figure 35: PRODORIC: a database developed and published by the Bioinformatics Competence Center of Braunschweig and the Institute of Microbiology of the Technical University of Braunschweig. PRODORIC is an acronym for Prokaryotic database of gene regulation, and it describes a large number of controlled vocabularies to annotated information on the regulation of gene expression in prokaryotes.

There are four ways to access the PRODORIC database: submitting a query on the web site, browsing through the content via GBpro genome browser, exploring regulatory networks via ProdoNet graph-visualizing tool, and accessing web services using the SOAP interface [27].

The database is a relational structure focused on genomes, and it allows the modeling of several biological features and molecular interactions, including operons, promoters, and protein complexes as shown on the UML diagram in figure 36. The database is publicly available for download on the web site.

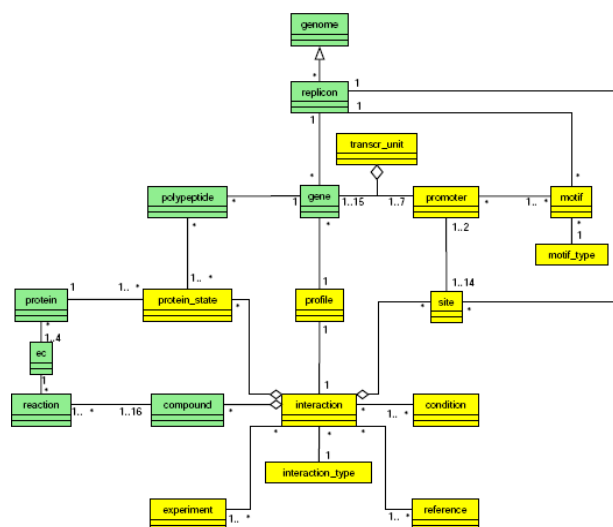


Figure 36: PRODORIC database's entity relation diagram, available for download on the database's web interface, where the user can integrate on their own program, performing local queries.

3.2.2. DBTBS

DBTBS is a reference database of transcriptional regulation for *Bacillus subtilis*, summarizing the experimentally-characterized transcription factors, their recognition sequence, and regulated genes [25].

The database was made via collection of experimental gene regulatory relations from published literature. DBTBS's current version contains a total of 120 transcription factors, 45 position specific scoring matrices, 1,475 promoters, 736 regulated operons, and 463 terminators [25].

The web interface of DBTBS's home screen, shown in Figure 37, supports the prediction of binding sites of transcription factors and the performing of queries of position weight of directly-inputted matrices or of PWMs created by the insertion of sequences.

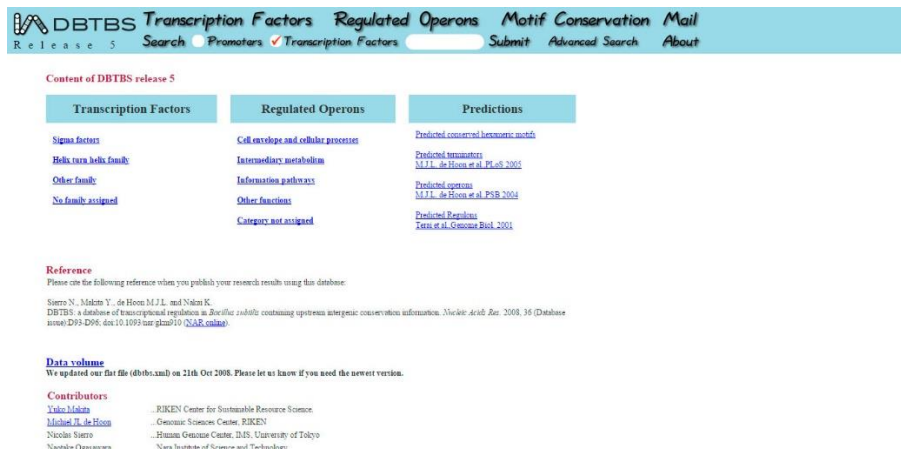


Figure 37: Database developed and published by the Human Genome Center of the Institute of Medical Science, University of Tokyo. It is a reference for *Bacillus subtilis* studies.

3.2.3. MtbRegList

This database is dedicated to the analysis of gene expression and regulation of the human pathogen *Mycobacterium tuberculosis*. The first release of the MtbRegList database contained a total of 315 annotated DNA motifs, divided in 72 transcription start sites, 119 promoters, 121 transcription factors binding sites, and 3 terminators, all data obtained from 56 researcher papers [78].

In addition, MtbRegList has a web interface in which the researcher can perform queries to their database, as shown in Figure 38; by selecting one gene of interest, the query shows gene products, gene or protein sequences, and positions in genomes.



Figure 38: MtbRegList's website: a database developed and published by the Département de Biologie (Biology department – in French) of the Université de Sherbrooke (University of Sherbrooke – in French). This database focuses on human pathogen *Mycobacterium tuberculosis*, analyzing gene expressions and regulation data.

There are two ways to make searches in MtbRegList: first, search requests of the whole genome information, besides the second type, which accepts queries for annotated DNA motifs, including transcription start sites and transcription factors binding sites. DNA motifs stored in the database are classified as either root patterns or DNA motifs. Each DNA motif stored comes from Root patterns, which are akin to consensus sequences and are obtained from literature and/or experimentally-identified DNA motifs. Also, the concept of signature is used to allow the user to query information at several motifs per request [78].

All information stored in MtbRegList comes from TubercuList [79]. Furthermore, the web interface provides links to COG from GenBank's annotation [80]. The front-end gives the opportunity to the user to download results in XML format or tab limited text format. Moreover, the interface presents a genome browser where the user can navigate through the genome choosing a specific genome region or center the genome browser in one gene of interest. MtbRegList was developed by the Département de Biologie (Biology department – in French) of the Université de Sherbrooke (University of Sherbrooke – in French) and it is accessible at <http://mtbreglist.genap.ca/MtbRegList/www/index.php>.

3.2.4. RegTransBase

RegTransBase is an open-access platform with a user-friendly interface. Its main goal is to cover a wide microbial diversity and provide a collection of experimental data to use in external computational tools [81].

RegTransBase's database is filled with the use of techniques of controlled vocabulary to capture knowledge in published scientific literature, describing a great number of regulatory interactions and containing several types of experimental data [81].

Another two tools were developed in association with RegTransBase, RegPredict, and RegPrecise. RegPredict is a web tool for reconstruction of transcriptional regulations in closely related prokaryotic genomes [82]. The second tool, RegPrecise, is a database to capture, visualize, and analyze transcription factor regulations that were reconstructed by RegTransBase [83].

This database contains information of 666 bacterial species of 224 genera, giving access to more than 19,000 experiments collected from 7,200 published papers [81].

As with the previous explained tools, RegTransBase also presents a web interface that covers six classifications that encompasses every aspect of the database. Three out of six categories present descriptions of genomes studied in relevant experiments. Two categories refer to experimental methodology and the goals of experiments, and the last category, the effector, uses a tree-like hierarchy in which classes are mainly based on MESH's chemical and drug categories [81]. A screenshot of the RegTransBase home screen is presented in Figure 39.

Browse by Experiments					
Genome			Experiment		Miscellaneous
Taxonomy	Relevance	Phenotypes	Technique	Result	Effector
<ul style="list-style-type: none"> Archaea (508) Bacteria (15195) 	<ul style="list-style-type: none"> Acetone production (42) Agricultural (1115) Amino acids production (1143) Animal Pathogen (809) Antibiotics production (489) more relevance (39) 	<ul style="list-style-type: none"> Acidophile (102) Alpha-hemolytic (155) Amylase production (5) Bacteriolytic (4) Barophile (0) more phenotypes (60) 	<ul style="list-style-type: none"> Binding analysis (3571) Function analysis (889) Genetic methods (8400) Other (1436) Protein analysis (5501) more technique (1) 	<ul style="list-style-type: none"> Gene/operon activation (5675) Gene/operon repression (3154) Operon structure characterization (1927) Plasmid replication (17) Exon/intron mapping (4126) more result (3) 	<ul style="list-style-type: none"> Amino Acids (266) Antibiotics (199) Carbohydrates (676) Coenzymes (5) Complex Molecules (34) more effector (8)

Figure 39: RegTransBase: is a database developed and published by the laboratories Genomics Division, Lawrence Berkeley National Laboratory, The Virtual Institute of Microbial Stress and Survival, and the Research and Training Center on Bioinformatics. The project is supported by the US Department of Energy Genomics GTL and by the Howard Hughes Medical Institute. The platform is open-accessed with a user-friendly interface; its main goal is to cover a wide microbial diversity and provide a collection of experimental data.

The database is accessible for free at <http://regtransbase.lbl.gov/cgi-bin/regtransbase?page=main>, where users can perform several biological searches, using gene names, effector names, and a full text of an abstract. They can download the database in the dump format, compatible with MySQL DMBS [81]. RegTransBase is developed and published by the laboratories Genomics Division, Lawrence Berkeley National Laboratory, The Virtual Institute of Microbial Stress and Survival, and the Research and Training Center on Bioinformatics. The project is supported by the US Department of Energy Genomics GTL and the Howard Hughes Medical Institute.

3.2.5. TRANSFAC

TRANSFAC (<http://www.biobase.de>) is a commercial database developed and published by BIOBASE. It is a database of transcription factors, their binding sites, nucleotide distribution matrices, and regulated genes [84].

TRANSFAC's data is complemented by another database, TRANSCompel [84] [85], which emphasizes the key role of specific interactions with transcription factors binding to their target sites, providing specific features of gene regulation on a particular cellular content.

TRANSFAC focuses on eukaryotic organisms, human, mouse, *Arabidopsis thaliana*, *Drosophila melanogaster*, and *Saccharomyces cerevisiae* [84]. This database is presented here because it provides a similar analysis in the field of gene regulatory networks.

This database is also linked to other databases and tools for the complementation of biological information; for humans, mouse, and rat genes, the database is respectively linked to HGNC, MGI, and RGD. It is manually updated by curators that search for suitable data in literature. Then, these data are inserted into a relational database via client input with the use of controlled vocabulary and of several automated functions [84].

TRANSFAC's web interface integrates versions of Match, Patch, and P-Match, providing great functionalities to TRANSFAC's front-end. Both tools are used to intensify the accuracy of binding site predictions. Match uses matrix-based transcription site binding searches and Patch uses pattern-based transcription site binding search [86] [87].

No image from TRANSFAC could be taken. The system is open only for costumers, and there is a free trial version for TRANSFAC available for those who request a login by filling a form at their website (<http://www.gene-regulation.com/pub/databases.html>).

3.2.6. PePPER

PePPER is a web interface for MolgenRegDB; this database is a collection of data for transcription factors, binding sites, and regulons for *Lactococcus lactis* [88] that can be accessed at <http://pepper.molgenrug.nl/>, and was developed by the Department of Molecular Genetics of the University of Groningen.

This web service was developed to mine regulons and transcription factors binding sites in any sequenced bacterial genome. Extending the database to published gene regularity networks of *Lactococcus lactis* and in addition for analysis, data from RegulonDB and DBTBS also are included on PePPER's web service [88].

All regulations and binding sites inserted in the database went through a process of confirmation via computational methods in which position weight matrices were calculated for all published TFBS. Tests were made using an intergenic area concatenated with the first 20 base pairs of their genes in order to search for DNA motifs, resulting in motifs of 6 to 18 base pairs. A database of all intergenic areas of *Lactococcus lactis* MG1363 was used as a background model. Afterwards, a manual test was performed comparing the resultant TFBS of the computational process with the published, data and samples that matched were added to the database [88].

PePPER's interface, shown in Figure 40, also presents a tool for TFBS prediction. Using conserved sequences at position -10/-35, the tool is capable of predicting TFBS for any bacterial genome, but also the tool may present different predictions for gram-negative or gram-positive bacteria [88].

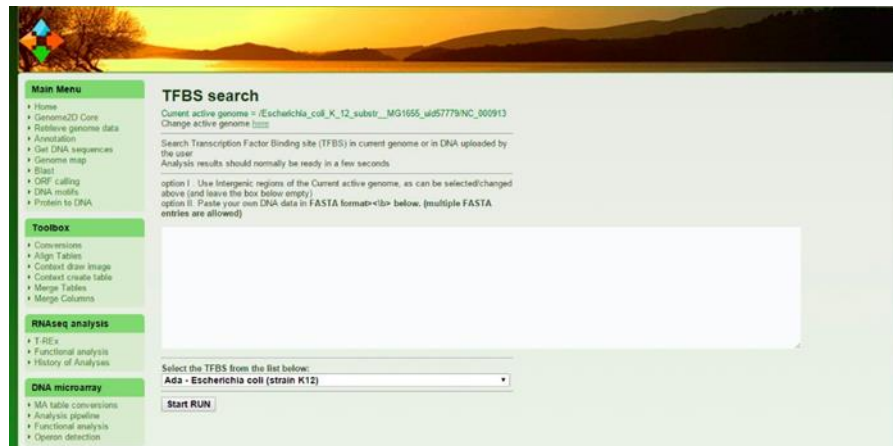


Figure 40: PePPER: a web service developed and published by the Department of Molecular Genetics of the University of Groningen. Figure showing the prediction tool of the Transcription Factor Binding Site.

3.2.7. Tractor DB

Tractor DB is a relational database in which computational predictions of binding sites of transcription factors for gamma-proteobacterial genomes are stored. It stores predictions of regulatory networks for 30 bacterial genomes using a weight matrix-based approach [89] [90].

The start point of the regulatory network predictions is the already published data of *Escherichia coli* K12, whose transcription factors of *E. coli* were used to create a statistical model, limiting the predictions of phylogenetic-related bacteria to *E. coli* [90].

Figure 41 presents Tractor DB's front-end, which allows the user to navigate through the regulatory interactions within a given regulon with a map that contains all known transcription factors and the regulatory interaction that interconnects them. The interface also allows the user to download the prediction information for every organism [90].

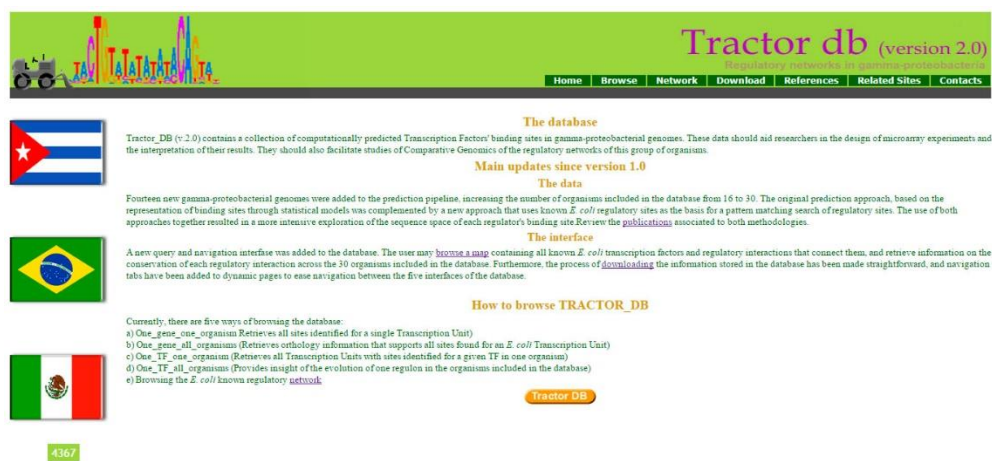


Figure 41: Tractor DB’s web interface. A computational prediction database of TFBS and regulatory networks based on the *Escherichia coli* regulatory network.

3.2.8. SwissREGULON

This database contains genome-wide annotations of regulatory sites in the intergenic regions of genomes; it contains regulatory site annotations for 18 prokaryotic genomes [91].

The database was populated with already-published data, but it also started with validated data. Some models were created to predict regulatory networks in other bacterial genomes using the tool MotEvo, a suite for TFBS prediction from multiple alignments of phylogenetic-related genomes [91][92].

The SwissREGULON database, shown in Figure 42, can be accessed at <http://swissregulon.unibas.ch/fcgi/sr>. The user can perform searches on any of the 18 genomes offered using GBrowser, a genome browser tool, to navigate through the genome and through their presented features. The web interface also presents a download section, with links to flat-files that contain information of regulatory predictions and of weight matrices used for the predictions [91].



Figure 42: SwissREGULON: A database for genome-wide annotations of regulatory sites in the intergenic regions of genomes.

3.2.9. Summary

Here, on this section, related works are summarized, comparing, by a compact view, the databases' contents and the analyses of the features. Are added databases RegulonDB [24] and MicrobesOnline [26] already discussed previously on the section 2.2 Related works for TAXI.

3.2.9.1. The content of the databases

In their databases, the related platforms store the following organisms:

- RegulonDB: 1 organism - *Escherichia coli K12 mg1655*;
- MicrobesOnLine: 3707 organisms - 1752 Bacteria, 94 Achaea, 119 eukaryotes;
- PRODORIC: 29 organisms - *Escherichia coli*, *Bacillus subtilis*, *Corynebacterium glutamicum*, *Pseudomonas aeruginosa* and more;
- DBTBS: 1 organism - *Bacillus subtilis*;
- MtbRegList: 1 organism - *Mycobacterium tuberculosis H37Rv*;
- RegTransBase: 658 organisms - *Bacillus subtilis*, *Corynebacterium ammoniagenes*, *Corynebacterium diphtheriae* and more;

- TRANSFAC: *Homo sapiens* (human), *Mus musculus* (mouse), *Arabidopsis thaliana*, *Drosophila melanogaster* (fruit fly), and *Saccharomyces cerevisiae* (yeast);
- PePPER: 1 organism - *Lactococcus lactis*;
- Tractor DB: 30 organisms - *Escherichia coli* K12, *Salmonella typhi*, *Erwinia carotovora*, *Photobacterium profundum* and more;
- SwissREGULON: 20 organisms – 17 prokaryotic, 3 eukaryotic – *Homo Sapiens*, *Mus Musculus*, *Saccharomyces cerevisiae*, *Escherichia coli*, *Bacillus subtilis* and more.

3.2.9.2. Data analysis features

Table 8 compares and summarizes the main data analysis and visualization of related platforms in which we base our aims for keeping our database updated in relation to new biological demands and preparing the system for the future.

Table 8: Summary and analysis of the features presented by all systems considered here

Feature	Genome browser	Network visualization	Network analysis	Raw data access	Data exchange	BM prediction
RegulonDB	+	+		+		
MicrobesOnline	+			+		
PRODORIC	+	+	+			+
DBTBS	+					+
MtbRegList	+			+		+
RegTransBase	+		+	+		+
TRANSFAC			+	+		+
PEPPER				+		+
TractorDB	+	+	+	+		
SWISSREGULON	+	+	+	+		

3.3. Methods

This section describes the methodology used to update the already well-known system, CoryneRegNet, going through the steps taken to update it, but also not forgetting the previous version of the system, comparing the updated or replaced concepts and addressing the update process to house new biological concepts and computational features.

3.3.1. Data integration

As previously discussed in section 2.3.2, that talks about the data integration for the TAXI database, biological databases have to deal with data stored in different ways in different databases; each group develops their point of view about the biological information and stores it in the way that best fits their needs.

For CoryneRegNet, the data source, which is the basis for the database, is formed by a collection of flat files, in which genome annotations, predicted operons, gene regulatory networks, and the membership of transcription factors to their families can be found.

In order to supply new biological demands, CoryneRegNet's update was made with modifications on the system architecture, and a totally new database was conceived to perfectly suit new paradigms addressed by the update.

Next two sections, 3.3.1.1 System architecture and 3.3.1.2 Data structure, will be addressing the update made on the system and trace a comparison line to previous CoryneRegNet's versions.

3.3.2. System architecture

CoryneRegNet's system architecture was developed as a publically available web-based software; as mentioned previously, the back-end has to cover different data sources of biological information, considering the different nuances of source files and databases.

Genome annotations and sequences were downloaded from NCBI in the GenBank format, and were implemented into the CoryneRegNet database with relevant data of the

gene regulatory network, imported to the database, derived from literature (inserted as PubMed link).

The system version is still divided into two main parts, as the former versions: a back-end with a software to parse the biological information and create the data structure described on the next section, and a user-friendly front-end to introduce all validated and predicted biological data stored in the CoryneRegNet database to the user. Moreover, a further description is made on the next sections. In figure 43 is shown a diagram of the system.

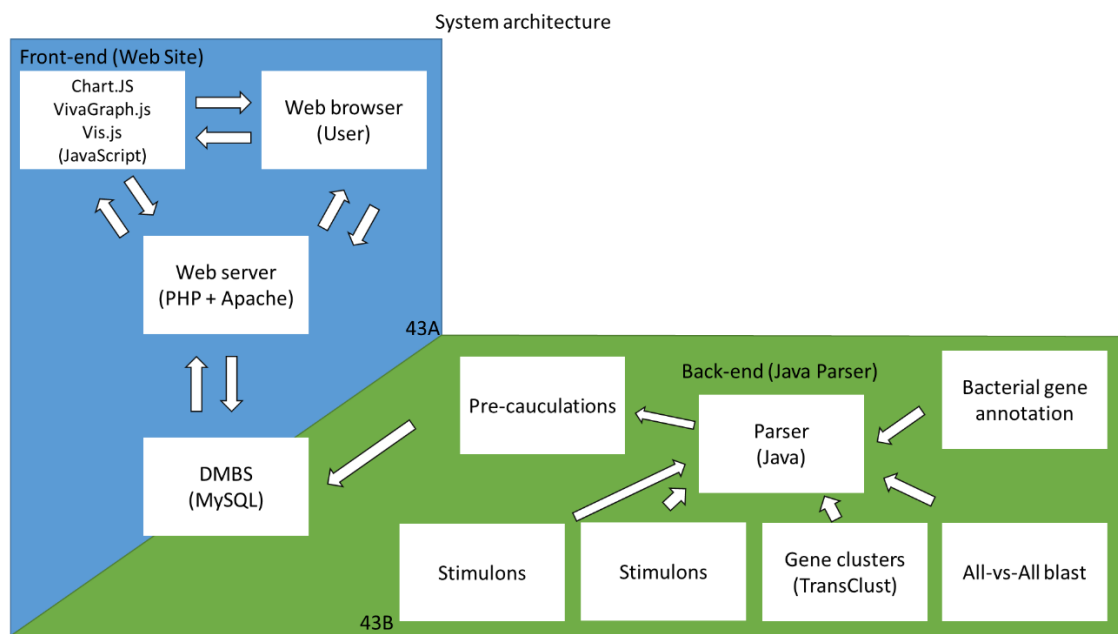


Figure 43: Figure displaying the system structure of the CoryneRegNet system. The architecture is divided in two parts: a back-end database that contains all results of precalculations of biding motifs and regulations transfer alongside genome annotations on 43B; and the front-end presentation on part 43A, which presents the interconnection between the actors used to create CoryneRegNet’s web interface.

The new parser software performs the same steps of the previous version, but on this updated version, all the steps of this import process are automatically performed, without user intervention, preventing errors on data handling or overturns on CoryneRegNet’s import pipeline. Replacing the previous version where the user has to execute the import process step by step in a pre-defined order to complete the process of creation of database.

The starting point of this import process is the creation of a basic database, inserting all genome annotations and sequences, already creating HMMer profiles used to predict new biding sites for transferring the regulations between organisms; a further explanation

about the use of HMMer for binding site prediction is given in section 3.3.2.

This important process also includes a BLAST search all-*vs*-all calculation using an e-value (expected number of higher scoring hits in random sequences) threshold of 10^{-10} for proteins; this calculation is used for homology detections, a process discussed in section 3.3.3. Afterwards, the process of network transfer is performed. To transfer the gene regulatory network from source organisms to others, a step that also has incorporated modifications from the previous pipeline version, and that is discussed on the section 3.3.4, is necessary.

The update of the new interface started with a new and modern layout, with new features and tools, bringing CoryneRegNet to a new level. The front-end was still developed using PHP and JavaScript as programming languages, running in HTML + CSS web pages for navigation.

The PHP programming language (*www.php.net*) is used to perform queries on CoryneRegNet's back-end database and populate tables in the front-end with the resultant data. On the other hand, the database was developed on MySQL (*www.mysql.com*) DBMS, following a new data structure in replacement of the former data structure of previous CoryneRegNet's versions. Additional information about new database structure is given at the following section.

Besides PHP, the development of the web interface also used JavaScript, this user-sided programming language is used to create graphs to visualize regulatory networks and statistical graphics. For graphs, two libraries were used: VivaGraphJS (<https://github.com/anvaka/VivaGraphJS>) and vis.js (*www.visjs.org*), libraries with different features and focused to best fit the desired graph. For statistical graphics, ChartJS library was used, a fast library to drawing graphics. Supplementary information about CoryneRegNet's web interface is given in section 3.3.5.

3.3.3. Data structure

As previously discussed, CoryneRegNet also went through an update process to create a new data structure to solve a demand that appeared over time. As new organisms were imported into the database, a considerable amount of data was generated, slowing the entire down.

The new data structure was developed with the use of the same ideals of the older data structure, where the database was based on ontologies, a data structure consisting of concepts that are linked through relations. The integrated data can be considered as a set of structured and named concepts, whereas the data sources are called controlled vocabularies [93].

In addition, the process of data import into the data structure suffered further modifications; starting with the insertions of genome annotations, all additional information were inserted in a dependently, centering the information of the organism as the major actor on the database.

The former ontology-based data structure, shown in Figure 44, is a structure divided into two main parts: the generalized data structure (GDS) and the ontology-based data structure; the second part is responsible for storing the main part of the database with the most important tables, which stores the biological concepts and the relations between them. The generalized data structure is a supporting set of tables in which are stored additional information of concepts and relations.

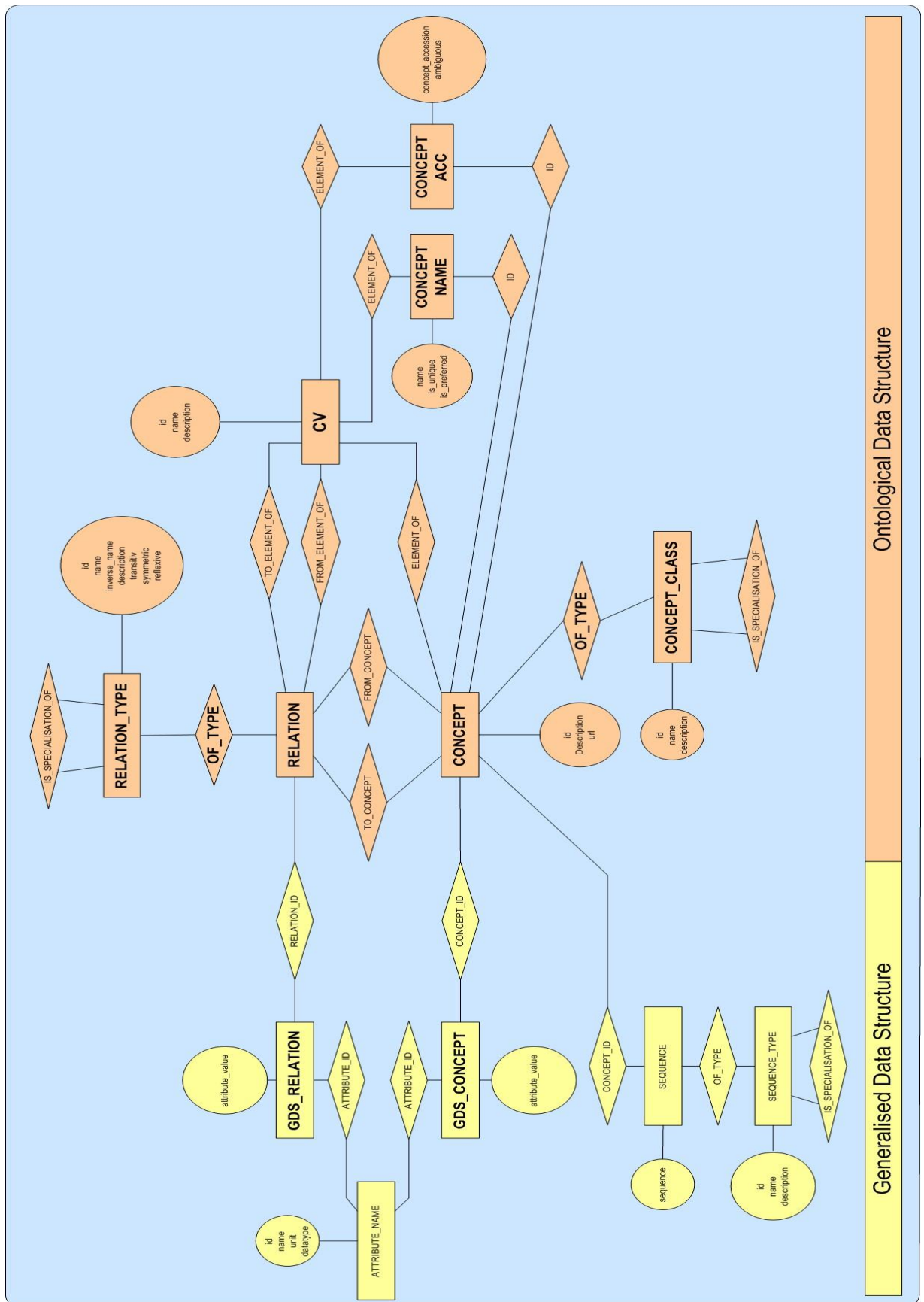


Figure 44: Ontology-based data structure of CoryneRegNet’s entity relation diagram. The structure is divided into two parts: Ontology-based Data Structure and Generalized Data Structure. The first part is the main part of the database, storing the main tables with biological concepts. The second part houses attached information to the biological concepts and relations [94].

With the growth of the biological information imported into the CoryneRegNet database with the insertion of new organisms of interest, the amount of data has expanded exponentially, generating a problem for the front-end queries, with a considerable increase of the database response time. The new database structure came to solve that problem, giving a new level of reliability and speed to the system.

The new database structure, developed for CoryneRegNet, shares the same ideals of the old database structure; starting with biological concepts, there were generated a set of entities to cover all concepts and their relations.

Creating tables for each biological concept allowed the division of the most important tables in many small tables, decreasing the quantity of rows per table and allowing the system to increase the total of studied organisms.

A table was generated for each biological concept. Organisms, genes, transcription units, or regulation units are examples of biological concepts with new-dedicated tables; also, relationships between concepts were stored in dedicated tables, decreasing the quantity of rows per table as well. The generalized data structure was also inserted into the new tables, avoiding the need of distinct tables for attached information. Figure 45 presents the entity relation diagram for CoryneRegNet's new data structure.

3.3.4. Binding site predictions

The prediction of binding sites, previously executed by CoryneRegNet in its back and front-end, were performed with the use of a tool based on the Position specific scoring matrix approach called PoSSuMSearch; a high PSSM-score in some region of a sequence often indicates a possible biological relationship of this sequence to the family or motif, characterized by the PSSM [95].

To maintain a good level of satisfaction for CoryneRegNet users, this former tool was replaced by a more accurate tool to generate a better binding site prediction on the web interface and back-end parser.

This new tool uses the technique of the profile Hidden Markov Models to predict a binding site on the DNA strand of a desired gene. This technique represents an important advance of the sensitivity of a sequence search for remote homology. It provides a probabilistic framework for sequence comparison and improve the detection of homologues [96].

The prediction of binding sites with HMMer is completed after the execution of several steps; the starting point of the transcription is the alignment of all validated binding sites over a valid input file for HMMer to create the HMM profile of the binding motif.

This sequence alignment is completed with the use of another computational tool, ClustalO. The Clustal Omega software is a tool that also uses the Hidden Markov Models technique to align a set of sequences. This new tool compares other multi-sequence aligners to deliver the same accuracy for smaller datasets, but for larger data sets, Clustal Omega outperforms other programs in terms of quality and execution time [97].

After the sequence alignment in the multifasta format, the resulting file was used to create the HMM, using the tool *HMMerbuild* of the HMMer package. Default parameters were used for this profile HMM creation, to have no influence of different parameters on the creation of several profiles for different transcription factors and organisms.

The last step of the binding site prediction is the prediction itself; to search the sites, the *nhmmer* tool, from the HMMer package, was used. This tool is specially designed for searching binding sites on DNA sequences. By comparing the Profile HMM against the

upstream sequence of a determined gene, it is possible to predict the binding site location.

As mentioned previously, the HMMer package is used in CoryneRegNet's web interface and back-end; for both predictions, the same steps are performed maintaining the best results for the users of the system.

3.3.5. Homology detection

With the massive amount of data generated by the process of DNA sequencing, biological studies to determine the function of proteins gained importance for biological researches, since the validations made with validated methods on laboratories are very expansive. To satisfy that role, computational methods to detect homology between proteins were developed.

For CoryneRegNet, a computational algorithm is used to cluster proteins into same groups across different organisms used on the study. These clusters were necessary so the transfer pipeline might detect regulatory behaviors of coupled proteins.

Previous versions of CoryneRegNet's pipeline used a clustering tool called FORCE, a tool that is motivated by a physically-inspired, force-based graph layout algorithm developed by Fruchterman and Reingold [98]. Where nodes from sub-graphs with high-weighted vertex connections should be arranged nearby, and nodes with low weight connectivity should be arranged far away, creating a layout to define the clusters using Euclidian single-linkage clusters.

To improve the reliability and speed of CoryneRegNet's back-end pipeline, the FORCE tool was also replaced by another program to complete the clustering step of back-end pipeline, maintaining the same quality of the results as the previous tool, TransClust, which was introduced in the pipeline.

TransClust executes the clustering of homologues based on Weighted Transitive Graph Projection; the main idea of this technique is to transform a given intransitive graph into a transitive one, adding or removing edges from the graph [99].

To create homologues clusters using TransClust, the BLAST all vs. all result is necessary, a process that is executed for all proteins present in studied organisms; with the BLAST result, a threshold value is determined to control the size of the clusters. For the new CoryneRegNet's version, a threshold for TransClust was used to generate results

similar to the previous used tool, maintaining the same quality of the results reported before.

3.3.6. Network transfer pipeline

The network transfer pipeline is the last step of CoryneRegNet's pipeline, predicting, on a target, the presence of a regulation based on a validated regulation of a source organism. This transfer pipeline requires the genome annotation parsed in the database and the detection of homologues genes across all studied organisms.

A transcriptional regulation is based in three major actors: the transcription factor, the binding motif, and the target gene. The transcription factor is the protein responsible for activating or repressing a gene transcription, attracting or repelling the RNA polymerase for the DNA sequence. The binding motif is the DNA sequence recognized by the transcription factor, upstream of the transcription start site; the position of the recognized sequence determines the activation or repression of the transcription initiation. The last actor of a transcriptional regulation is the target gene, the gene in which the transcription begins or is repressed by the transcription factor.

The process of network transferring from one organism to another depends directly of homologues clustering, as far as the regulation is predicted based on the genes present on the target organism that are homologues to genes on source organisms.

Starting with a list of transcription regulations from the source organism, the network transfer pipeline, using the homologues clusters, detects the presence of similar regulatory networks on target organisms.

For each regulation on the list of transcription regulations, the pipeline tests the presence of an orthologue transcription factor in the target organism; if the presence of the TF is confirmed in the target organism, the pipeline senses the presence of the orthologue target gene. After the construction of binaries of orthologue genes, the transfer pipeline starts the prediction of the binding site. If all transcription regulation actors were detected on a target organism, the regulation is successfully transferred from one source organism to a target organism.

3.3.7. Visualization

The CoryneRegNet's web interface was developed accordingly with the proposal of presenting all biological information stored in the database, facilitating then the access to information. A new user-friendly interface was developed, replacing the old interface with a more accessible layout and using new tools for a better presentation of the biological nuances of organisms housed in the database.

The update of the user interface redesigned the layout of the web page, introducing a new, more user-friendly way of front-end navigation, replacing old tools for faster and more reliable ones. That was accomplished by introducing on the new front-end, HMMer with HMMerLogo for binding sites prediction, Chart.js for creating statistical graphics, and VivaGraph for drawing larger graphs, and for small ones, Vis.js.

3.3.7.1. User interface

The new CoryneRegNet's web interface was developed to attend new demands of the system; accordingly, the web interface had changes in the access to the database for faster navigation and implementation of new tools.

As mentioned previously, for the back-end and also for the interface, the binding site tool was replaced. The former PoSSuM search tool was replaced by the HMMer package, introducing a new quality for the binding site prediction.

The old tool for graph visualization, GraphViz, developed as a Java applet, no longer works in modern web browsers, making impossible the visualization of regulatory network graphs; two JavaScript graph drawers, executing the graph draw in different situations, replace it. VivaGraph is the first and more powerful tool used for graph drawing, used to create of bigger graphs with more complexity and larger quantity of nodes. Vis.js, the second tool, was used to draw smaller graphs for a couple of nodes with reduced quantity of edges connecting them.

The graphic-drawing library, introduced in new CoryneRegNet's interface to replace the old library, Chart.js, creates interactive graphics that are easier to understand, giving more usability to statistical data analyses.

When the user visits CoryneRegNet, its home page presents a welcome message to

the user and basic information about the system, links for documentation, links to download the back-end software and the current available database. Figure 46 presents a screenshot of CoryneRegNet's welcome page.

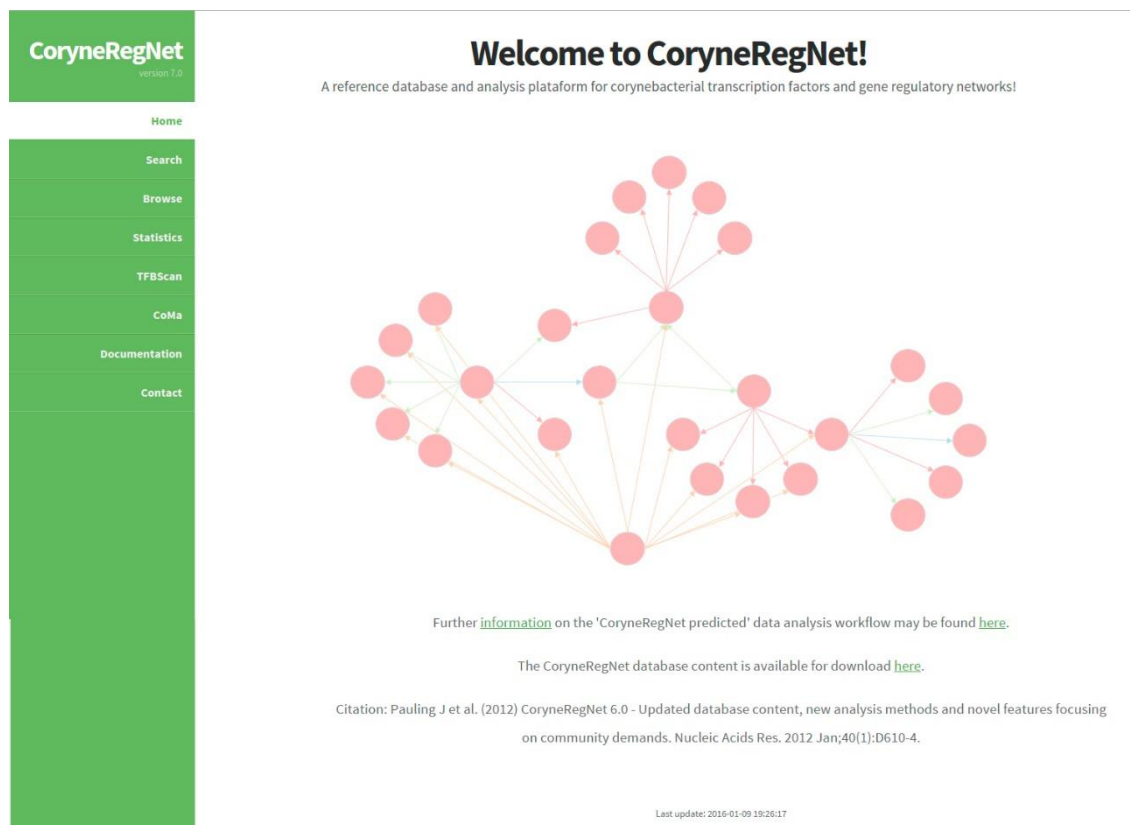


Figure 46: CoryneRegNet 7.0's intro page. On the intro page, the user can find the menu to navigate on the entire CoryneRegNet's web interface, performing searches and analyses of genes, proteins, transcription units, or gene regulatory networks. On the intro page, the user has access to documentation and download links of CoryneRegNet's back-end software and database, and to the last database update.

The intro page already presents the navigation menu of the web interface to the user on the left side of the page; on the menu, the user can access CoryneRegNet's major sections and perform analyses. Search, Browse, Statistics, TFBSscan, and CoMa are the major parts, followed by documentation and contact. Figure 47 shows CoryneRegNet's menu in details. A list of all major parts of the system is described next.

- Search – On the search section, the user can perform searches on the database, query for a biological term that the user can search for genes, proteins, regulator type, or modules.

- Browse – The user can navigate through all organisms stored on the database, gaining access to specific organism pages with statistics for the organism, and access to genes and to the regulatory network graph.
- Statistics – This section presents all statistics of the database or divide it specifically for each organism.
- TFBSscan – In this section, the user can enter binding sites to create a pattern and search for a bacterial genome inserted on the database, or enter an upstream sequence and use all transcription factors of one organism to predict a binding site on the entered sequence.
- CoMa – Contradiction on microarrays; the user can search for contradictions on microarray data in the database and confirm regulatory networks.

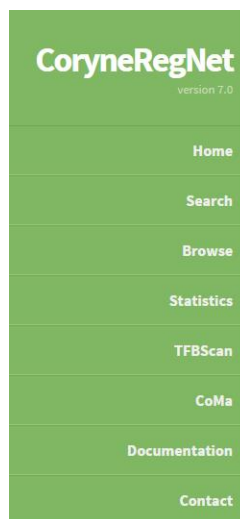


Figure 47: A detailed CoryneRegNet’s screenshot, showing the major sections of the front-end in which the user can perform analyses and gain access to biological information stored in the database. The user can search for terms in the database, browse through genomes stored in the database, access statistics, and compare it between organisms, predict binding sites on upstream sequences of genes, and verify contradictions on microarray data.

In CoryneRegNet’s search section, the user can perform a search for a biological term that points to a gene, protein, regulator type, or gene module. For gene, the user can use the gene identifier, gene name, or an alternative gene identification. For proteins, protein identification or protein name.

The search process can be executed on all organisms stored in the database or on one specific organism. Also, on the search procedure, the user can specify the type of term used or make search on all types, increasing the time required by the database to respond

to the query. Also, the results can be sorted by type of term, organizing the result at the user's will. Figure 48 shows the search form.

Search

Figure 48: Search page in which the user can perform searches of biological terms in the CoryneRegNet database, searching for a specific organism or for all organisms, for a special type or all types, and sorting the result by a specific type of term.

The result of a search is presented in another page, respecting the specified options on the search page. Figure 49 presents a search result for the term “*cg0444*” with no specification of type, searching in all organisms and sorted by gene identification.

Search

Used parameters							
Organism	all						
Search	cg0444						
In field	all						
Short by	geneID						

Result							
1 entry found							
GeneID	Alt. GeneID	ProteinID	Protein name	Regulator type	Predicted operon	Organism/Genome/Plasmid	Regulation
cg0444	NCgl0358	YP_224669.1	transcriptional regulator, MerR family	HTH_3		Corynebacterium glutamicum ATCC 13032 (NC_006958)	Click!

Figure 49: Result of a search in the CoryneRegNet database: a search performed using term “*cg0444*” on all organisms, in any field, and sorted by gene identification.

The results page is divided into two tables: the first table presents the parameters used on the search in which the user can confirm the parameters. The second and most important table presents the result; in this example case, just one entry in the database

was found. The result presented: gene and protein identifications, regulator type, predicted operon, organism in which the entry came from, and a regulation link if the entry is inserted in one.

CoryneRegNet's browse section interface allows the user to access directly each organism inserted in the database, with a link to the organism page. The statistics page, shown in Figure 50, presents full statistics of the database divided into two subsections: main and specific statistics.

In the main statistics subsection, part 50A, two tables are presented: the general statistics of the database divided per elements such as genes, proteins, regulations, binding motifs, and clusters, and a second table also with links for each page of the organisms, where the user can find specific statistics for each organism.

Part 50B presents links for the database statistics; the user can access quantity of regulations, quantity of regulator families, distribution of transcription factors, distribution of co-regulations, comparison between regulations and co-regulations, and distribution of transcription sites distances. Each statistic is calculated for all organisms in the database and specifically for each organism.

Main statistics			
Elements	Quantity	Organisms/Genomes/Plasmids	Genes
Organism	12	Corynebacterium jeikeium K411 (CR931997)	2104
Genes	30466	Corynebacterium glutamicum R (NC_009342)	3052
Proteins	30465	Corynebacterium pseudotuberculosis 1002 (CP001809)	2057
Regulations from wet lab	9380	Corynebacterium urealyticum DSM 7109 (AM942444)	2024
Predicted regulations	3971	Corynebacterium kroppenstedtii DSM 44385 (NC_012704)	2018
Regulators from wet lab	314	Corynebacterium diphtheriae NCTC 13129 (NC_002935)	2272
Predicted regulators	332	Corynebacterium glutamicum ATCC 13032 (NC_006958)	3058
Regulated genes from wet lab	5987	Escherichia coli K12 (NC_000913)	4237
Predicted regulated genes	2672	Corynebacterium pseudotuberculosis C231 (CP001829)	2053
Biding motifs from wet lab	7636	Corynebacterium aurimucosum ATCC 700975 (NC_012590)	2531
Predicted binding motifs	2943	Corynebacterium efficiens YS-314 (NC_004369)	2950
Hidden Markov Models	243	Corynebacterium pseudotuberculosis FRC41 (NC_014329)	2110
Clusters	3776		

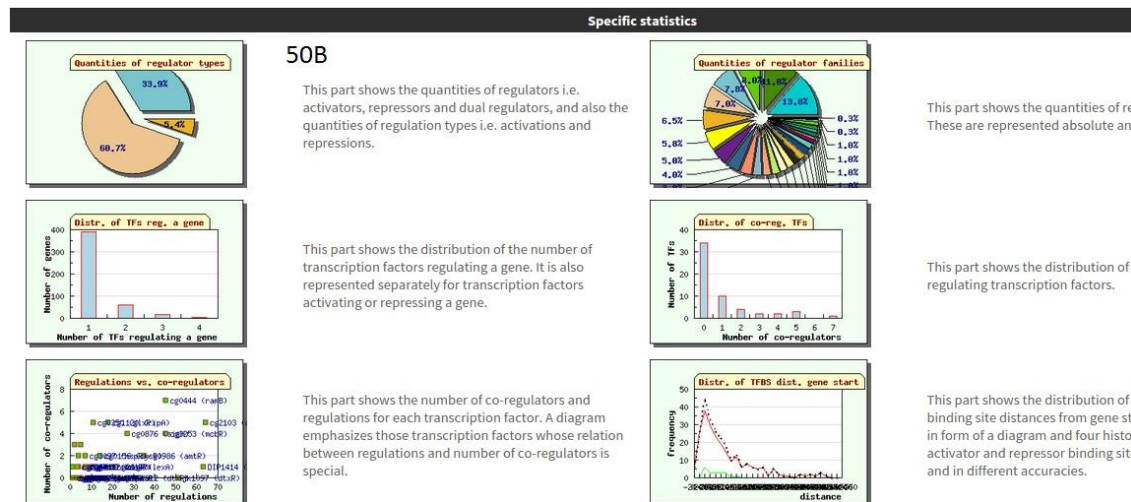


Figure 50: Statistics page in CoryneRegNet’s interface, divided into two parts: 50A, page of main statistics, general information about the database and links for pages of organisms for each organism, and part 50B, which shows further specific statistics of the regulatory networks stored in the database.

In the transcription factor binding scan (TFBScan) section, the user can predict binding sites on sequences deposited in the database or use transcription factor binding motifs created during CoryneRegNet’s back-end run to predict binding sites on a desired sequence. Figure 51 presents the two types of binding site predictions.

Starting with part 51A, the binding prediction is executed on the subsection with the use the binding motifs generated by CoryneRegNet’s pipeline, using binding motifs of transcription factors of the database. The user enters a maximum amount of 10 sequences, with a 1,000-character length each, and selects a list of binding motifs of all organisms or one exact organism to be used in the prediction. In addition, the user can control the quality of the HMMer profile used, the p-Value cut off, and if the prediction will be executed on both strands.

The second part, 51B, represents the prediction of binding sites over housed sequences in the database using an HMMer profile, obtained from sequences entered by the user. Also, the user may control search parameters, like both strands and target genes inside operons. Also controlled by the user, the prediction is executed on a target organism and cannot be executed on all organisms.

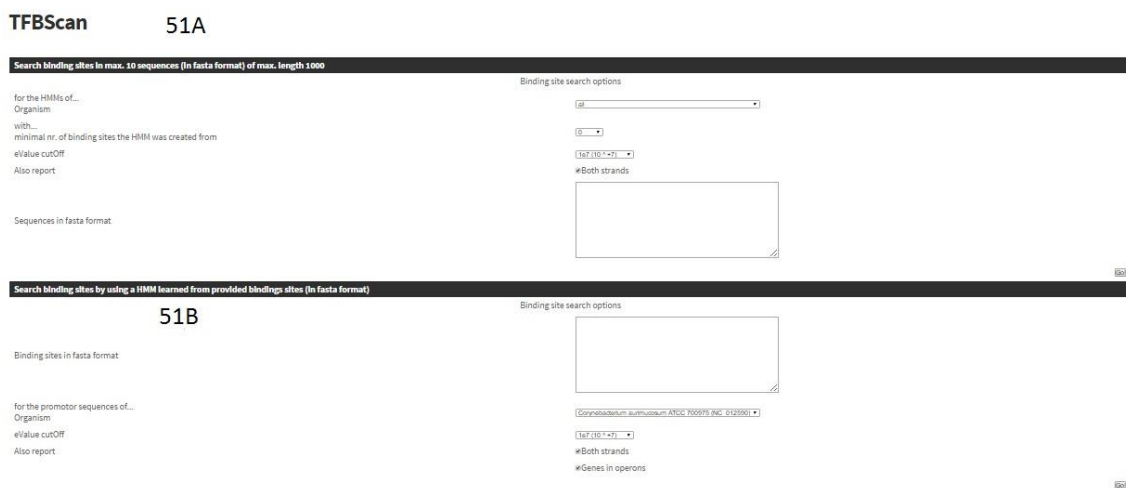


Figure 51: Transcription factor binding scan. The CoryneRegNet’s user can perform binding site predictions using HMMer profiles stored on the database or create their own HMMer profile and run the prediction on the genome sequences in the database.

Executing the prediction of the binding sites on part 51A using three sequences, shown on Table 9, with the validated binding sequence for gene “*cg0444*” inserted inside the sequences “AATACTTTGCAAA,” and the HMMer profiles generated for bacterial genome *Corynebacterium glutamicu* ATCC 13032, the result of the prediction is presented in Figure 52.

Table 9: Sequences used for the binding prediction of the figure 3.19A, predicting binding motifs in entered sequences with the insertion of the binding site “AATACTTTGCAAA” of gene the “*cg0444*.”

Identification	Sequence
01	CATGCTAGCTAGCTACGAATACTTTGCAAAATCGATCGACTAGCTGATCGA
02	AATACTTTGCAAATCGATCGACTGACTGACTAGCTAGCTACG
03	CTAGCTAGCTAGCTACGATCGAATACTTTGCAAA

Binding Motif prediction result

Matches										
Using HMM profiles of <i>Corynebacterium glutamicum</i> ATCC 13032 (NC_008950)										
Source Gene ID	Source gene name	Target Gene ID	Target Gene ID	Predicted operon	Reverse Complement	eValue	Bias	Bits	Sequence	Candidates for homologous proteins in validated original list
cg0317		01	-		f	0.4	2.6	-0.1	AAAGTAT	
cg0317		02	-		f	0.4	2.6	-0.1	AAAGTAT	
cg0317		03	-		f	0.4	2.6	-0.1	AAAGTAT	
cg0444		01	-			0.00026	1.9	11.8	AATACTTTGCAA	
cg0444		02	-			0.00026	1.9	11.8	AATACTTTGCAA	
cg0444		03	-			0.00026	1.9	11.8	AATACTTTGCAA	
cg0444		01	-		f	0.0072	1.9	7.6	TTTTCGAA	
cg0444		02	-		f	0.016	1.9	6.9	TTTTCGAA	
cg0444		03	-		f	0.019	0.9	6.3	TTTTCGAA	
cg0908	amrR	02	-			0.014	0.0	6.6	CGATCGACTG	
cg0908	amrR	01	-			0.043	0.0	5.9	ATCGATCG	
cg0908	amrR	01	-		f	0.053	0.4	5.0	CGATCGA	
cg1704		01	-		f	0.4	2.6	-0.1	AAAGTAT	
cg1704		02	-		f	0.4	2.6	-0.1	AAAGTAT	
cg1704		03	-		f	0.4	2.6	-0.1	AAAGTAT	
cg1114	livA	03	-			0.13	0.6	3.6	TCGATTA	
cg2000	rgpR3	01	-			0.094	1.2	3.9	GAATACTTTGC	
cg2000	rgpR3	03	-			0.094	1.2	3.9	GAATACTTTGC	
cg2000	rgpR3	02	-			0.33	1.8	3.4	AATACTTTGC	

Figure 52: Result of binding predictions executed by inserting three sequences, with the validated binding motif of the “cg0444” gene inside the three sequences. The nhmmer tool from HMMer package was able to find the correct binding site inside all inserted sequences.

Executing the binding prediction in Figure 51B, inserting a set of sequences to generate an HMMer profile, using sequence “TAGACCATACGGTCTA” ten times repeated and as a background model, the genome of *Corynebacterium glutamicum* ATCC 13032. The result page is presented in Figure 53, containing minor changes in comparison with Figure 52. The result page does not show the information of the source gene and whether the regulation already exists in the database, in reference to the fact that the binding motif might not exist in the CoryneRegNet database.

Binding Motif prediction result

Matches										
Used as background model <i>Corynebacterium glutamicum</i> ATCC 13032 (NC_008950)										
Target Gene ID	Target gene name	Predicted operon	Reverse Complement	eValue	Bias	Bits	Sequence			
cg0054	-			0.0013	0.2	24.5	AGACATACGGTCTA			
cg0059	-			0.46	0.4	13.3	CATACGGTCT			
Genes on operon										
cg0059	-	OP_cg0059								
cg0059	-	cg0059 cg0072								
cg0059	-	OP_cg0059	f	0.46	0.4	13.3	CCATACGGTC			
Genes on operon										
cg0131	-	cg0413 cg0451								
cg0131	-			1.9	0.1	11.4	TACGGTCTA			
cg0131	-		f	1.9	0.1	11.4	ATACGGTCT			
cg01454	-			2	0.2	11.3	CCATACGGT			
cg0201	rne	OP_cg0201		3.7	1.1	10.5	AGACCATAC			
Genes on operon										
cg0201	-	cg0201 (rnl) cg0205 (rplU) cg0206 (rpmA)								
cg0201	-	OP_cg0201	f	3.7	1.1	10.5	TAGACCATA			
Genes on operon										
cg1103	-	cg0209 cg0209								
cg1103	-		f	9	0.1	9.3	GTATGCCA			
cg1103	-		f	9	0.1	9.3	GTATGCCA			
cg1103	-			9	0.1	9.3	ATACGGTC			
cg1144	-			9	0.1	9.3	ATACGGTC			
cg1144	-		f	9	0.1	9.3	CATACGGT			
cg1451	-		f	9.3	0.1	9.2	CATACGGT			
cg2451	-	OP_cg2451								
Genes on operon										
cg2451	-	cg2451 cg2450 cg1399 (pik) cg2388 (pikC) cg0387 cg0399								
Alternative operons										
cg0443	-	OP_cg0443								
cg0908	-	OP_cg0908	f	9.3	0.1	9.2	TACGGTCT			
cg0908	-	OP_cg0908	f	9.3	0.1	9.2	TACGGTCT			
Genes on operon										
cg0908	-	cg0908 cg0907 cg0906								
Alternative operons										
cg0908	-	OP_cg0908								
cg0908	-	OP_cg0908	f	9.3	0.1	9.2	ATGCCAGA			
Genes on operon										
cg0908	-	cg0908 cg0907 cg0906 cg0905								
Alternative operons										
cg0908	-	OP_cg0908								

Figure 53: Screenshot of the result of the binding motif prediction shown on figure 51B. A DNA sequence repeated ten times was used to generate the HMMer profile to search the motif on the complete genome sequence of *Corynebacterium glutamicum* ATCC 13032.

The last major section of CoryneRegNet’s interface is the contradictions on microarrays. This feature of the interface gives to the front-end the ability to check the consistence of microarray data with known regulatory networks.

The analysis of a contradiction in a microarray experiment, using known regulatory networks stored in the database, can be performed using three types of data entrance: copy and paste of the data into a text field, uploading to the front-end a tab-delimited flat file, or using stimulons data already stored in the CoryneRegNet database. This is shown in Figure 54.

CoMa

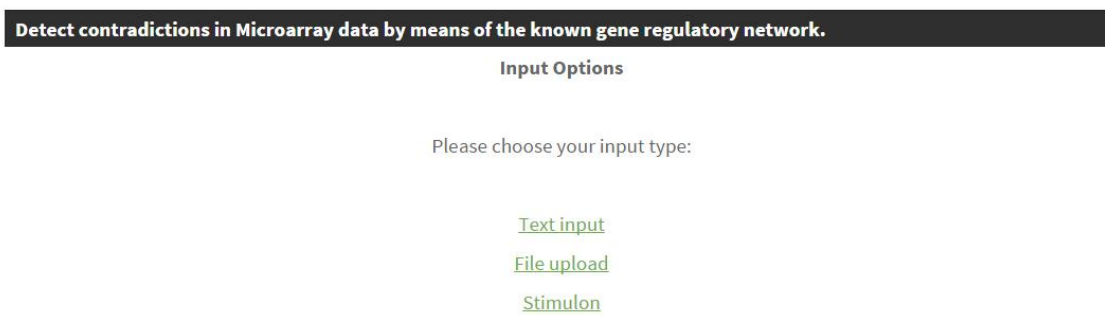


Figure 54: Contradictions on microarrays. A screenshot showing the input options of contradictions on microarrays, presenting the three ways of analyzing the contradictions in the context of gene regulatory networks stored in the database.

By using one of the three types of data input to query the contradictions on microarrays data, the web interface queries the database for regulatory network information and compares de input data with the query result.

The calculation of contradictions consists in comparing if a gene is up-stimulated or down-stimulated by another gene, with the regulatory information for that pair of genes. In other words, if “*genA*” is up-regulated by “*genB*” in micro array data, but on the context of regulatory network, “*genA*” is repressed by “*genB*,” then the interface calculates one contradiction on the information. Executing one example of contradiction calculation, the information in Table 10 is used as input data. The result page is shown in Figure 55.

Table 10: Microarray data used for a toy test on the contradictions on microarrays of CoryneRegNet’s interface.

Gene identifier	M-value
cg0444	1.9
cg0445	-1.8
cg0446	1.8
cg0447	-2.5
cg0448	-1.7
cg2831	-1.6

The result page shown in Figure 55 presents the calculation of contradictions of microarray data with regulatory network data housed in the CoryneRegNet database. The result shows five putative contradictions, two being for “*cg0444*,” two for “*cg0446*,” and the last one for “*cg2831*.”

This CoryneRegNet’s feature could provide hints for incorrect operon prediction, missing regulatory interactions and putative inconsistencies in the experimental setup.

CoMa

Putative regulatory contradictions					
GeneID	mValue	Predicted operon	Contradictory regulations	Non contradictory regulations	No further putative explanations?
cg0444	1.9		Gene: +, but repressor cg0444 : + Gene: +, but activator cg2831 : -	Repressor cg0350 (*) Activator cg2092 (*)	
cg0446	1.8	OP_cg0445	Gene: +, but repressor cg0444 : + Gene: +, but activator cg2831 : -	Repressor cg0350 (*) Repressor cg1120 (*) Activator cg2103 (*)	
cg2831	-1.6		Gene: -, but repressor cg2831 : -	Repressor cg2115 (*) Activator cg2092 (*)	

(+) upregulation or activation
 (-) downregulation or repression
 (*) mValue insignificant or GeneID is not in the Microarray.

Putative contradictions in operons	
Predicted operon	Genes with significant mValue
OP_cg0445	cg0445 , mValue: -1.8 cg0446 , mValue: 1.8 cg0447 , mValue: -2.5

No regulatory contradictions			
GeneID	mValue	Predicted operon	Regulations
cg0445	-1.8	OP_cg0445	Repressor cg0350 (*) Repressor cg0444 , mValue: 1.9 Repressor cg1120 (*) Activator cg2103 (*) Activator cg2831 , mValue: -1.6 Activator cg2092 (*)
cg0447	-2.5	OP_cg0445	Repressor cg0350 (*) Repressor cg0444 , mValue: 1.9 Repressor cg1120 (*) Activator cg2103 (*) Activator cg2831 , mValue: -1.6
cg0448	-1.7		Repressor cg0350 (*) Repressor cg0444 , mValue: 1.9 Repressor cg1120 (*) Activator cg2103 (*) Activator cg2831 , mValue: -1.6 Activator cg2092 (*)

(*) mValue insignificant or GeneID is not in the Microarray.

Unknown regulations or GeneID not found in database		
GeneID	mValue	Predicted operon

No contradictions in operons	
Predicted operon	Genes with significant mValues

Figure 55: Contradictions on microarray results page, showing comparisons of regulatory network information with microarray data. The results were achieved using a toy test present on table 10, in which five contradictions were found, considering two auto-regulations on the calculation. This feature of CoryneRegNet could provide a background for operon predictions, elucidating errors on the microarray data or missing regulatory interactions.

To conclude the analysis of CoryneRegNet’s major features, the biological concepts, which contain specific web pages to explain the biological information regarding the concepts, will now be discussed.

The first biological concept presented is the organism. The web page contains specific statistics of the organism, and the next figures are sections of the organism page. Figure 56 presents the first section of the organism web page, with main statistics for the organism.

Organism: *Corynebacterium glutamicum* ATCC 13032

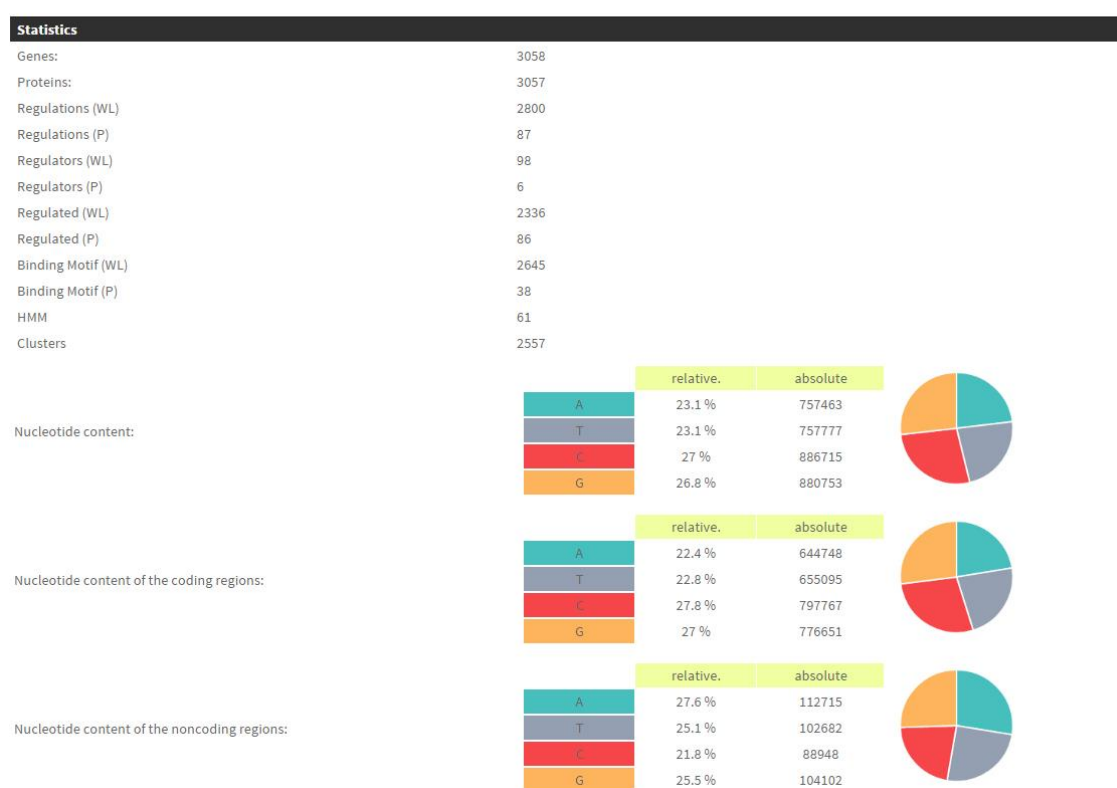


Figure 56: Screenshot of the first section of the organism web page for the bacteria *Corynebacterium glutamicum* ATCC 13032. In this section, the main statistics of the organism are presented as quantity of genes, proteins, transcription units, regulations, and three graphics of nucleotide contents.

This first section contains general information about the organism, the quantity of genes, proteins, transcription units, regulatory networks, HMMer profiles, and clusters. This section also presents graphics related with the nucleotide content of the entire organism, but also divided by coding and noncoding regions of the organism. The next section, presented in Figure 57, is optional and not shown for each organism, so not all of

them have data of modules and stimulons.

8 Modules:		
Module name		Number of genes
CRN-module Carbohydrate Metabolism (NC_006958)		266
CRN-module Cell Division and Septation (NC_006958)		12
CRN-module Cellular Program (NC_006958)		1
CRN-module Macroelement and Metal Homeostasis (NC_006958)		204
CRN-module Regulation of cell wall biosynthesis (NC_006958)		2
CRN-module Sigma Factor Module (NC_006958)		4
CRN-module SOS and Stress Response (NC_006958)		189
CRN-module Specific Biosynthesis Pathways (NC_006958)		129

10 Stimulons:		
Stimulon name	Stimulon description	Number of genes
mitomycin c	wt + mitomycin c vs. wt, minimal medium	81
delta_dtxr	delta_dtxr vs. wt, complex medium + fe	257
delta_ssur	wt vs. delta_ssur, minimal medium + cysteine	31
delta_ltbr	delta_ltbr vs. wt, complex medium	50
propionate	wt cgxii acetate/propionate vs. wt cgxii acetate	160
transition phase	res167 transition vs. res167 exponential, minimal medium mm1	111
delta_sigm	delta_sigm vs. wt, minimal medium mm1	37
delta_mcbr	delta_mcbr vs. wt, minimal medium	134
delta_lexa	delta_lexa vs. wt, minimal medium	350
delta_zur	delta_zur vs. wt, minimal medium	26

Figure 57: Second section of the organism web page. This section is optional and not shown for all organisms. Here, one organism can present modules, stimulons, both, or none.

In this second section, two optional tables are presented. In the first table, the module table, links for modules that belong to the organism are presented, and on the second table, a list of stimulons is presented, with links for each stimulon page and information related to the microarray data stored in CoryneRegNet used on contradictions on microarrays.

The last section of the organism page, shown in Figure 58, presents all genes with related information and links to subsequent pages. Also, at the bottom of this section, the link for the entire gene regulatory network of the organism is presented.

Organism has: 3058 Genes

GeneID	Alt. GeneID	Gene name	ProteinID	Protein name	Regulator type	Predicted operon
cg0001	NCgl0001	dnaA	YP_224295.1	CHROMOSOMAL REPLICATION INITIATOR PROTEIN	DnaA	
cg0002			YP_224296.1	hypothetical protein predicted by Glimmer		
cg0004	NCgl0002	dnaN	YP_224297.1	DNA POLYMERASE III, BETA SUBUNIT		OP_cg0004 (Pri.)
cg0005	NCgl0003	recF	YP_224298.1	DNA REPAIR AND GENETIC RECOMBINATION PROTEIN		OP_cg0004 (Pri.)
cg0006	NCgl0004		YP_224299.1	hypothetical protein		OP_cg0004 (Pri.)
cg0007	NCgl0005	gyrB	YP_224300.1	DNA GYRASE SUBUNIT B		OP_cg0004 (Pri.)
cg0008	NCgl0006		YP_224301.1	hypothetical protein		
cg0009	NCgl0007		YP_224302.1	UNCHARACTERIZED MEMBRANE PROTEIN		OP_cg0010 (Pri.)
cg0010	NCgl0008		YP_224303.1	hypothetical protein		OP_cg0010 (Pri.)
cg0012	NCgl0009		YP_224304.1	PUTATIVE TRANSCRIPTION REGULATOR PROTEIN	ROK	
cg0013	NCgl0010		YP_224305.1	Bacterial regulatory proteins, tetR family		OP_cg0014 (Pri.)
cg0014	NCgl0011		YP_224306.1	Helix-turn-helix protein, copG family		OP_cg0014 (Pri.)
cg0015	NCgl0012	gyrA	YP_224307.1	DNA TOPOISOMERASE (ATP-HYDROLYSING)		OP_cg0015 (Pri.)

•
•
•

Graph visualization

[Go!](#)

Figure 58: Last section of the organism page, showing all genes for that organism with links for gene and operon pages. Also, at the bottom of the page, there is a link for the graph of the entire gene regulatory network for the referred organism.

The last section of the organism page is filled with all genes that belong to the organism; the gene table contains major information about the genes with links for gene page for each gene and for predicted operons in which the genes are inserted. How a gene can be predicted in more than one operon is explained in links for each operon in which the gene is inserted with an identification related with if the operon is a primary or a secondary operon, a new feature presented in this CoryneRegNet's update.

The gene concept web page will be explained, also divided in sections, on the next figures, emphasizing the major features of the concept page, main information about the gene, and protein encoded by the gene, prediction of binding sites on the upstream sequence of the gene, and attributes including an optional sequence logo for transcription regulators.

The web pages for genes are automatically generated, querying in the database for all information regarding the gene, creating external links when necessary and internal links with other concepts related to the gene. The first section of gene concept page is shown in Figure 59.

Gene: cg0444

Main informations			
Gene			
Gene	cg0444		
Gene Name			
External link	NCBI		
Protein			
Protein	YP_224669.1		
Protein name	transcriptional regulator, MerR family		
External link	NCBI		
Protein class	Transcription Factor		
Stimulated by 3 stimulons			
Stimulon	Stimulation type	M-value	PubMed
delta_dtxr01	-	-1.05	16469103
delta_lexA01	-	-1.11	
delta_mcbR01	+	0.8639384504239717	15853877
Organism			
Organism	Corynebacterium glutamicum ATCC 13032 (NC_006958)		

Figure 59: The first section of gene concept web page containing basic information for the gene, encoded protein of the organism, and optional stimulon information that is presented if the gene is up-stimulated or down-stimulated.

This first section presents main information about the gene, such as gene and encoded protein identification, and optional links for external databases, such as NCBI or RegulonDB. Also, optionally, links for microarray stimulon data are presented, which regulate the gene up or down with the m-Value for stimulation and necessary PubMed link regarding to the reference of the stimulon. The last subsection of the gene concept page is the link for the organism to which the gene belongs.

The second section of gene concept page presents the attributes of the gene, such as the first and last base pair position of the gene, in which the codon start the translation of the gene starts, and which kind strand the gene it is: normal or complementary. Occasionally, it shows the regulator type of the gene: if the gene regulates itself or if the gene is a mutant. And the last information, optionally shown, is the HMMer logo, if the gene is a transcription factor. The second section shown on figure 60:



Figure 60: Attribute section of the gene concept page that presents information such as the first and last base pair, strand, and the HMMer logo, if the gene is a transcription factor.

The HMMer logo shown in Figure 60 is dynamically-generated querying the HMMer profile of the transcription factor from the database and send the result for a web service. Skyline is a web service that generates interactive sequence logos representing aligned sequences and the profile Hidden Markov Models [100].

Skyline responds to the query with a link to download the HMMer logo as a .png image file or a generated .json file with the information to generate an interactive “div” in the web page with the HMMer logo. On CoryneRegNet’s interface the second option was used, not requiring the generation of a local file for the HMMer logo.

The next section of the gene concept page regards to the regulations in which the gene is inserted; each table of this section is optional, since the gene does not have to be under any regulation; each table queries the regulation for the database automatically.

The regulations are subdivided in types. In total, there are eight possible combinations of regulation types, the type refers to if the regulation is validated by a wet lab or predicted by the system during the prediction pipeline, if the regulation is controlled by a sigma factor or by a transcription factor, and if the gene is controlling the regulation or being controlled by another gene. An example of this section is shown in Figure 61.

Regulated by 1 gene (Sigma Factor) (Validated)							
Gene ID	Gene name	Protein ID	Protein Name	Activator/Repressor	Predicted operon	More information	
cg2092	sigA	YP_226152.1	RNA POLYMERASE SIGMA 70 FACTOR	Activator	OP_cg2091 (Pri.)	click!	

Regulates 53 genes (Transcription Factor) (Validated)							
Regulated by 3 genes (Transcription Factor) (Validated)							
Gene ID	Gene name	Protein ID	Protein Name	Activator/Repressor	Predicted operon	More information	
cg0350		YP_224590.1	TRANSCRIPTIONAL REGULATOR, CRP/FNR FAMILY	Repressor	OP_cg0352 (Pri.)	click!	
cg0444		YP_224669.1	transcriptional regulator, MerR family	Repressor		click!	
cg2831		YP_226801.1	Bacterial regulatory protein, LuxR family	Activator		click!	

Regulated by 1 gene (Transcription Factor) (Predicted)							
--	--	--	--	--	--	--	--

Figure 61: Screenshot of an example of the regulation section of the gene concept page. This example is based on the regulations of the gene “cg0444” of the bacteria *Corynebacterium glutamicum* ATCC 13032. This gene encodes a transcription factor with 53 regulations validated in laboratory.

The example presented in Figure 61 presents only four types of regulation in which three tables are formed by regulations with biological validation and one with predicted information. Referring to protein type, three tables are for transcription factors and one for sigma factors, and the last subdivision shows three tables in which the gene is regulated and one table where the gene regulates other genes.

Also, on the gene concept page, it is possible to predict binding sites on the upstream sequence of the referred gene or if the gene is a transcription factor, and one can use the HMMer profile so the gene might predict binding sites on upstream sequences of other genes. Figure 62 presents the two types of binding predictions.

Binding site prediction (for this regulator in other upstream sequences)

For the promoter sequences of...
 organism:

Also report: Both strands
 Genes in operons
 eValue cutOff

[Go!](#)

Binding site prediction (in the upstream sequence of this)

For the HMM's of...
 organism:

With a minimal quality of:

Also report: Both strands
 eValue cutOff

[Go!](#)

Figure 62: Types of binding site predictions presented on the gene concept web page, where binding sites of transcription factors can be found on the upstream sequence of the current gene or if the gene encodes a transcription factor; the binding site for other genes can also be predicted.

As previously discussed about CoryneRegNet’s TFBSscan feature, the prediction process is similar. To predict binding sites on the upstream sequence of the current gene, one can choose to use either all available HMMer profiles or those from a specific organism. Also, the minimum quality of the HMMer profile can be specified, electing a

group of transcription factors, and the minimum value of p-Value cut off for limiting the result for a desired excellence can be regulated, and the user may choose if the result will be predicted for both strands or not.

To use the HMMer profile of the current gene, the parameters are modified a bit: the prediction must happen in one specific organism, in both strands or not, in genes inside operons, and is possible to control the p-Value cut off of the prediction.

The last section of the gene concept page presents the homologue candidates of the gene: homologue groups generated by CoryneRegNet's back-end pipeline during the homologue detection step. This section also presents the gene and amino acid sequences retrieved from the annotation file of the organism, and the bottom presents a link for the graph in which the current gene is inserted with a cut off line for gene layers. Figure 63 presents the screenshot of the last section.



Figure 63: The last section of the gene concept page with candidates for homologues of the current gene, sequences of gene and amino acid, and a link for the graph in which the current gene is inserted.

3.3.7.2. Network visualization

Gene regulatory visualization is a feature present on CoryneRegNet's web interface, where the user can access the visualization of specific genes or an entire gene regulatory network of an organism. The network visualization presents biologically validated data alongside predicted transferred information by CoryneRegNet's pipeline.

As shown on the previous section, the network visualization can be accessed from the organism concept page and gene concept page, but there is also a third way available to generate network visualization: the prediction generated by the web interface also can generate a network visualization.

The graphs generated for network visualization on CoryneRegNet's web interface can also perform some biological analyses, activating or deactivating a set of regulations, simulating the behavior of sigma factors on regulatory networks on bacterial cell.

The construction of a graph is performed by a JavaScript library, VivaGraphJS; this library is known by its fast graph-building with a certain level of interactivity, helping the user to generate a graph that fits to his/her needs.

In CoryneRegNet’s case, at the regulatory network graph, every node represents a gene, and the connection between genes represents a regulation, with different colors and line formats. Also, the node is a link that opens a modal box on the right side of the screen with specific information for clicked links. Figure 64 presents a basic generated graph for gene “*cg0444*” of bacteria *Corynebacterium glutamicum* ATCC 13032.

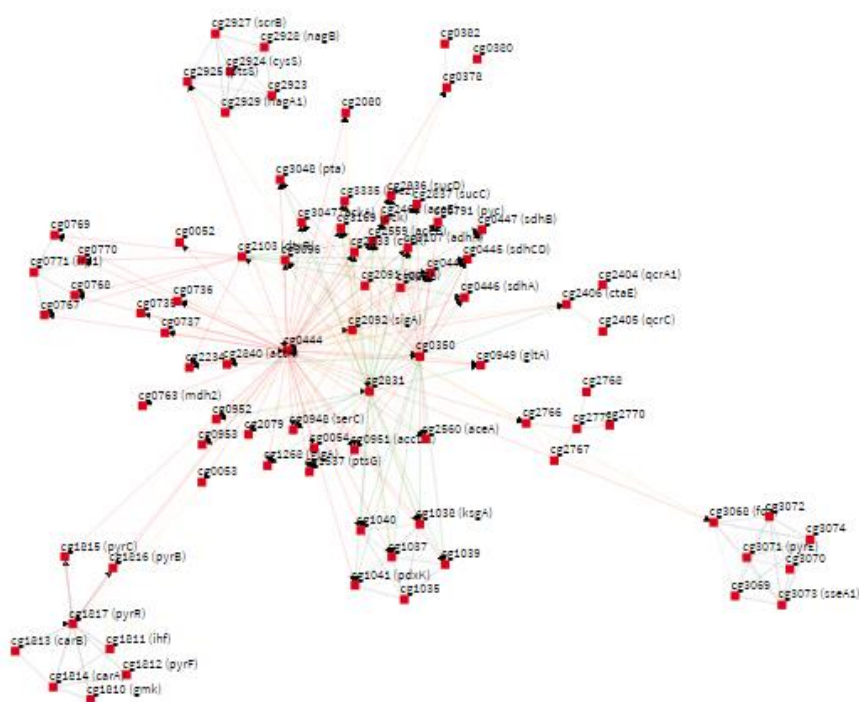


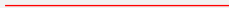



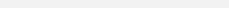

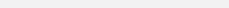

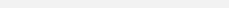


Figure 64: A graph generated for gene “*cg0444*” of *Corynebacterium glutamicum* ATCC 13032 with a depth cut-off of one, connecting only the first layers of the regulations with the gene. On the graph, the red nodes represent the genes and the colored lines represent the regulations that connect the genes or the genes inside operons.

Different colors and line shapes are used to represent the regulations between two genes, representing these kinds of regulation: activating, repressing, and both. The evidence of the regulation is represented by a straight line if the regulation is validated by laboratorial experiments or a dashed line if the regulation is predicted. Table 11 presents the colors and line formats used during the graph creation.

Table 11: Colors and line formats used to create the connection between genes on CoryneRegNet's graphs.

Line	Description
	Transcription factor activating the transcription of a target gene
	Transcription factor activating the transcription of a target gene in which the gene is not the first gene of the operon
	Transcription factor repressing the transcription of a target gene
	Transcription factor repressing the transcription of a target gene in which the gene is not the first gene of the operon
	Transcription factor activating and repressing the transcription of a target gene
	Transcription factor activating and repressing the transcription of a target gene in which the gene is not the first gene of the operon
	The source gene is a sigma factor
	Deactivated regulation
	Gene in operon
	Biologically validated regulation
	Predicted by a CoryneRegNet's regulation

A json file drives every graph generation that includes all genes of one organism, with the regulations, in which the genes were inserted in. This json file might be generated by a PHP script, depending of which type of graph will be generated.

The first type of graph shows the entire regulatory network of an organism. Figure 65 presents an example of this graph, using the gene regulatory network of *Corynebacterium glutamicum* ATCC 13032.

Graph

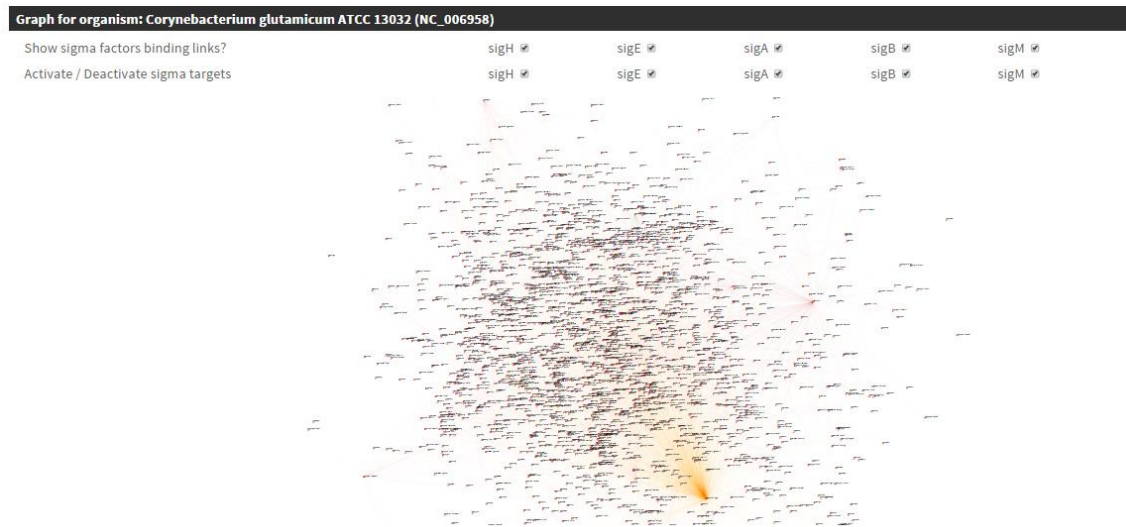


Figure 65: This screenshot presents the complete regulatory network of *Corynebacterium glutamicum* ATCC 13032 with all five sigma factors that regulates the transcription of genes in the bacterial organism.

Figure 65 also presents a feature that addresses the regulation control by the sigma factors present in the bacterial genome. There are two options for each sigma factor to adjust how the regulations will be displayed. The first option controls if the regulation between the sigma factor and the target gene will be displayed, simulating the absence of the sigma factor on the cell. The second option, also related with the regulation control of the sigma factor, turns this option off; all regulations directed to target genes, in which the same genes are also the target for a chosen sigma factor, are disabled, except for the sigma regulation. In other words, deactivating a chosen sigma factor only deactivates the driven regulations of the transcription factors, leaving the regulation of the chosen sigma factor active. This feature simulates the activation and deactivation of a sigma factor on the gene regulatory network.

Starting with the graph generated for the gene “*cg0012*” of *Corynebacterium glutamicum* ATCC 13032, Figure 66 presents, in more details, the use of the sigma factors’ regulatory network control.

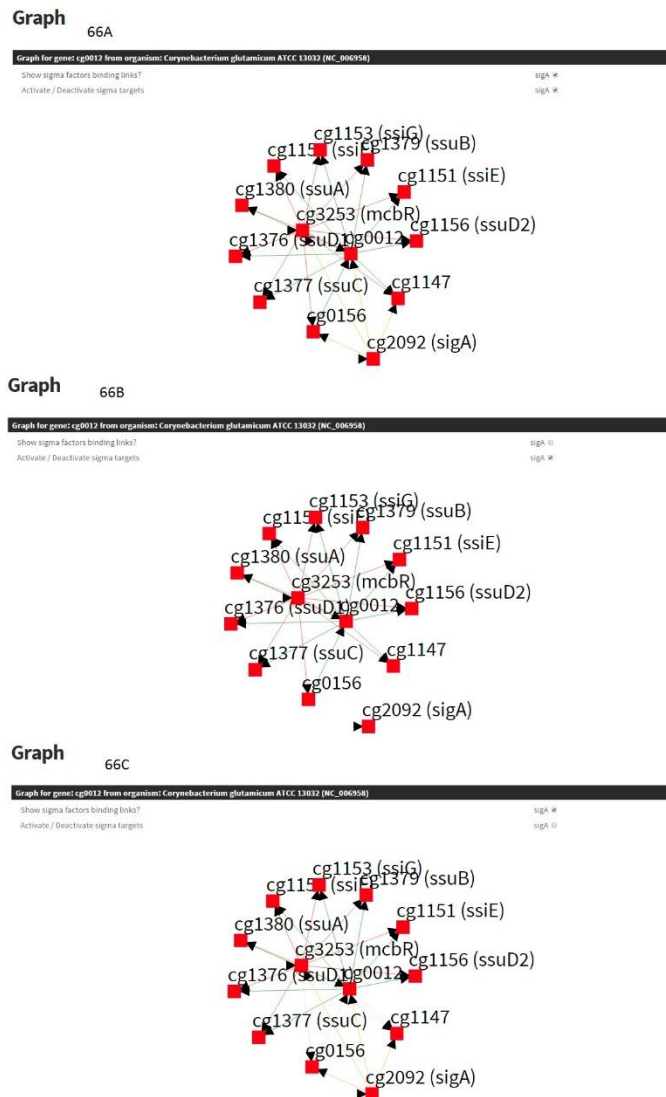


Figure 66: The screenshot shows all options available for simulating the control of sigma factors on the graph generated for gene “*cg0012*” of *Corynebacterium glutamicum* ATCC 13032. Figure 66A shows both options activated, simulating the presence and activation of the sigma factor “*cg2092 - sigA*” in the graph. Figure 66B simulates the absence of the sigma factor “*cg2092 - sigA*” in the graph, hiding the presence of regulations for that sigma factor. The last figure, 66C, shows the presence of the sigma factor “*cg2092 – sigA*,” but in this example, the regulation is deactivated graining the regulation targeting to the target genes of the sigma factor.

Another feature introduced by CoryneRegNet’s is a modal box. This box is shown on the right side of the screen when the user double-clicks on one gene; it presents specific information for the specific gene, with links for the gene concept page and transcription unit page. It also displays two mini graphs dividing the regulations in which the gene is inserted in biologically-validated regulations and predicted regulations. Figure 67 presents an example of a modal generated for the gene “*cg0012*”.

3.4. Results and Discussion

This section presents the results of the update of the CoryneRegNet system; a synthesis of CoryneRegNet's evolution across previous published versions will be presented, addressing the requirement of a new version of the system.

The discussion of CoryneRegNet's evolution is followed by the discussion of a new regulatory transfer pipeline, developed to cover a greater number of possible regulations and for greater accuracy; the new pipeline presents new features that were not explored on previous versions.

The last subsection presents a new regulatory network for three example organisms: *Corynebacterium glutamicum* ATCC 13032, *Escherichia coli* K-12 MG1655, and *Corynebacterium pseudotuberculosis* FRC41, presenting new biologically-validated information for the first two organisms, with results predictions made with the new transfer pipeline. For *C. pseudotuberculosis*, only the results for the new transfer pipeline are presented.

3.4.1. CoryneRegNet's evolution

The evolution of the CoryneRegNet system was driven by biological demands for comparison of gene regulatory networks between organisms of interest; new features were added to all versions to help researchers to access full biologically-validated gene regulatory networks or predictions based on regulations, with one or more organisms as source, with ease.

Based on only one organism as model organism, *Corynebacterium glutamicum* ATCC 13032, CoryneRegNet's first version was developed to describe the features present on these specific bacteria. Fifty-three transcription factors and 430 regulations are examples of statistics inserted in CoryneRegNet's first version [101].

In the second version, other three organisms related to *Corynebacterium glutamicum* and other three *Corynebacterium* were added. Also, the feature of transcription factor binding motif analyses was added. The number of transcription factors were raised to 64, with a total of 607 regulations [102].

In the third version, a new well-studied organism, *Escherichia coli* K-12 MG1655,

was added. Data were extracted from RegulonDB [24], a previously well-discussed database with the regulatory network of the referred bacteria. This version had a total of 213 TFs and 2,912 regulations [103].

Important features were added in CoryneRegNet's fourth version. For the first time, the researcher could perform analyses of contradictions on microarrays; as discussed already, the user could confirm or disapprove regulations on microarray data stored in the database or inserted by the user [104].

Also, in CoryneRegNet's fourth version, the feature of gene clusters was introduced, in which the genes were grouped by similarity, using the results of a BLAST all *versus* all genes and proteins. This release had the same amount of TFs and regulations as in the previous version, but stimulon data and gene clusters were added to it [104].

In CoryneRegNet's fifth release, which was an internal release, one of biggest features of the system was introduced; the first version of the transfer pipeline was released, performing a transfer of regulations of *Corynebacterium glutamicum* for all other *Corynebacterium* [94].

This major feature of the fifth version divided the database in experimental and predicted versions; the experimental database houses the biologically-validated data to PubMed ID for publications, when it is possible. The predicted database stores the results of the transfer pipeline performed by CoryneRegNet's back-end, generating a new database, separated for predicted data [94].

The 6.0 Version of CoryneRegNet is the consolidation of the transfer pipeline in which the prediction is performed for all stored organisms in CoryneRegNet, using *C. glutamicum* ATCC 13032 as a base organism, and all fully sequenced and annotated *corynebacteria* [94] are present in the database.

The summary of CoryneRegNet's evolution is presented in Figure 68, where the major features of each version are shown on a timeline with all versions. The amount of information is presented in Table 12, with a graphic of the image in Figure 69.

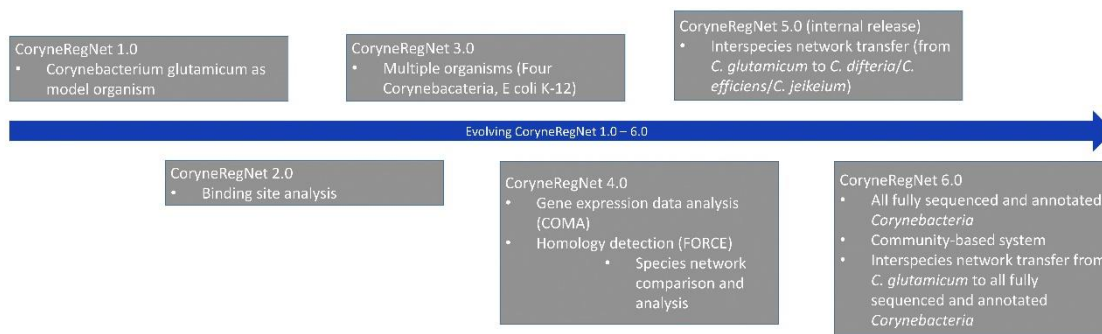


Figure 68: Evolution of the CoryneRegNet system of all versions, from 2006-2011. The major improvements per version are presented, such as the addition of new organisms with important features, such as COMA and gene clusters.

Table 12: Table with the amount of information added to the CoryneRegNet database per version, from the first version, with only one organism, to version 6.0, with 12 organisms.

Version	Organisms	TFs	Reg. genes	Regulations	BMs	PWMs	Stimulons	Clusters
1.0	1	53	331	430	192	23	–	–
2.0	4	64	499	607	274	29	–	–
3.0	5	213	1632	2912	1522	130	–	–
4.0	7	213	1632	2912	1522	130	8	4548
5.0e	11	245	1986	3712	1759	144	11	5421
5.0p	11	350	2888	4928	2553	249	11	5421
6.0e	12	245	1986	3712	1759	144	14	3719
6.0p	12	482	3946	6352	3429	381	14	3719

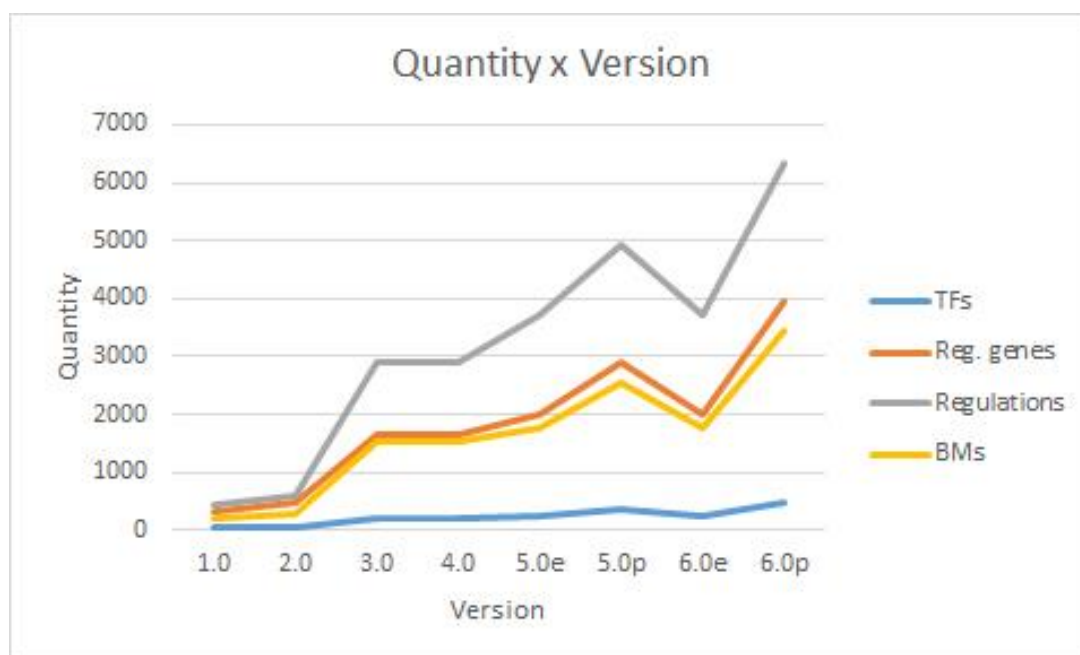


Figure 69: Graphic showing the evolution of quantities of transcription factors, regulated genes, regulations and binding motifs. The graphic shows the data growth of CoryneRegNet's versions.

Table 12 and the graphic in Figure 69 clearly show the exponential growth of stored data in the CoryneRegNet database; as previously discussed, this growth generated some problems for CoryneRegNet’s system. As data were added to the database, the use of the CRN front-end became a barrier to perform a good research. Therefore, to add new organisms, an update of the database was required and indispensable. Figure 70 represents this problem graphically.

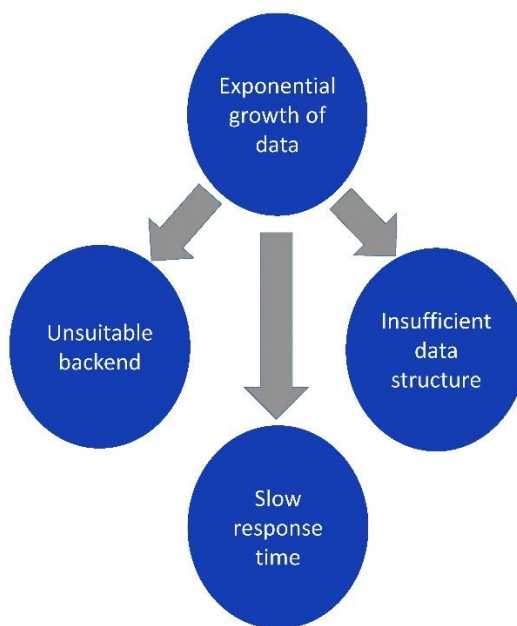


Figure 70: Graphic representing the major problem generated by CoryneRegNet’s evolution, in which the exponential growth of data made the back-end unsuitable, the data structure insufficient, and the time of response high.

CoryneRegNet’s 7.0 version discussed in this work, updated not only features or the front-end, but the source data also received new data from new analyses on regulatory networks of *Corynebacterium glutamicum* ATCC 13032 and *Escherichia coli* K-12 MG1655 [24]. Table 13 displays the amount of data present in CoryneRegNet’s 7.0 version.

Table 13: Amount of data stored in the CoryneRegNet database in version 7.0.

Version 7.0	Organisms	TFs	Reg. genes	Regulations	BMs	HMMs	Stimulons	Clusters
Validated	12	314	5987	9380	7636	243	14	3776
Predicted		332	2672	3971	2943			

While comparing Tables 12 and 13, a vital difference of experimental, validated data and predicted data can be seen in version 7.0. In the versions 5.0 and 6.0, the database was divided into two databases with distinct levels of information, one for experimental and another for predicted data. In version 7.0, the database was converged in only one database again, and the data was differentiated with flags in the table where the data were stored. This was a solution for the waste of hard disk space generated by storing the same data twice in CoryneRegNet's both levels of information.

Using only one database facilitates the general navigation on the CoryneRegNet front-end, as all data can be accessed with faster queries, not requiring comparisons across big tables with redundant data.

The update of the CoryneRegNet database facilitates the addition of organisms in the database, since every time an organism is added, data for genes, proteins, transcription units, homologues, gene clusters, validated regulation (when present), and predicted regulations are generated. That generates a vast amount of data for each added organism and, as discussed before, these data are stored in big and slow-to-access tables. Now the data is stored by being divided in more tables for faster queries, as shown in Table 14.

Table 14: Comparison of response times of the same queries performed in the CoryneRegNet databases versions 6.0 and 7.0.

Query	CRN v 6.0	CRN v 7.0
All genes	0.09s	0.08s
All genes of an organism	0.05s	0.01s
All predicted regulation units	0.25s	0.11s
All predicted regulation units with respective binding motif	0.33s	0.16s

As seen in Table 14, the response time of important queries decreased in the new database, making it possible to add new organisms without performance loss in the front-end, supporting the same quality of information presented before.

3.4.2. New network transfer pipeline

Continuing with CoryneRegNet's evolution, a new regulatory network transfer pipeline was developed also, designed to cover new nuances acquired during the time of development of the system.

The new pipeline added new features in comparison with older ones. When studies of regulatory networks for a couple of organisms appeared, the new pipeline used all organisms inserted on the database as a source of regulations.

CoryneRegNet's previous versions were using only *Corynebacterium glutamicum* ATCC 13032 as source organism, transferring all regulations inserted in that organism to all other 11 organisms in the database. In the 7.0 version, regulations from *Corynebacterium jeikeium* K411 or *Escherichia coli* K-12 MG1655, for example, are used as source regulations to make transfers for other organisms. Table 15 presents the organisms with the related amounts of regulations used on CoryneRegNet.

Table 15: Relation of organisms with the respective amounts of genes, transcription factors, target genes, regulations, and binding motifs.

Organism	Genes	TFs	Target genes	Regulations	BMs
<i>Corynebacterium jeikeium</i> K411	2104	1	51	51	21
<i>Corynebacterium glutamicum</i> R	3052	-	-	-	-
<i>Corynebacterium pseudotuberculosis</i> 1002	2057	-	-	-	-
<i>Corynebacterium urealyticum</i> DSM 7109	2024	-	-	-	-
<i>Corynebacterium kroppenstedtii</i> DSM 44385	2018	-	-	-	-
<i>Corynebacterium diphtheriae</i> NCTC 13129	2272	4	68	70	33
<i>Corynebacterium glutamicum</i> ATCC 13032	3057	98	2336	2800	2645
<i>Escherichia coli</i> K12 MG1655	4237	206	3486	6413	4917
<i>Corynebacterium pseudotuberculosis</i> C231	2053	-	-	-	-
<i>Corynebacterium aurimucosum</i> ATCC 700975	2531	-	-	-	-
<i>Corynebacterium efficiens</i> YS-314	2950	5	46	46	20
<i>Corynebacterium pseudotuberculosis</i> FRC41	2110	-	-	-	-
Total	30466	314	5987	9380	7636

Since more than one organism was used as source of regulations, some cases could be generated by the transfer pipeline. A biologically validated regulation can be predicted by the pipeline, using another related organism as source; in this case, the regulation is stored only as a complementation for the validated one. Table 16 presents an example of a validated regulation with the complementation of predicted regulations.

Table 16: An example of a validated regulation complemented with predicted regulations.

Organism	Source gene	Target Gene	Evidence	Source organism	Source gene (SO)	Target organism (SO)
<i>Corynebacterium jeikeium</i> K411	<i>Jk1097</i>	<i>Jk0315</i>	Validated	-	-	-
<i>Corynebacterium jeikeium</i> K411	<i>Jk1097</i>	<i>Jk0315</i>	Predicted	<i>Corynebacterium diphtheriae</i> NCTC 13129	<i>DIP1414</i>	<i>DIP0625</i>
<i>Corynebacterium jeikeium</i> K411	<i>Jk1097</i>	<i>Jk0315</i>	Predicted	<i>Corynebacterium glutamicum</i> ATCC 13032	<i>cg2103</i>	<i>cg0466</i>
<i>Corynebacterium jeikeium</i> K411	<i>Jk1097</i>	<i>Jk0315</i>	Predicted	<i>Corynebacterium glutamicum</i> ATCC 13032	<i>cg2103</i>	<i>cg0468</i>

Table 16 presents an example of a validated regulation: gene *jk0315* is being controlled by the transcription factor *jk1097* of the organism *Corynebacterium jeikeium* K411. The regulation was predicted three times using a regulation from two other organisms: *Corynebacterium diphtheriae* NCTC 13129, with one source regulation, and *Corynebacterium glutamicum* ATCC 13032, with two source regulations.

For predicted regulations, the same may happen, where a predicted regulation is predicted with the use of using an organism and an orthologue regulation as source. A second organism was used again to predict exactly the same predicted regulation. Again, the second prediction is only stored as a complementation of the first prediction. Table 17 presents an example of the predicted regulation being predicted by more than one source.

Table 17: An example of a predicted regulation complemented with other predicted regulations.

Organism	Source gene	Target Gene	Evidence	Source organism	Source gene (SO)	Target organism (SO)
<i>Corynebacterium efficiens</i> YS-314	<i>ce0948</i>	<i>ce1514</i>	Predicted	<i>Corynebacterium glutamicum</i> ATCC 13032	<i>cg2888</i>	<i>cg1568</i>
<i>Corynebacterium efficiens</i> YS-314	<i>ce2494</i>	<i>ce1514</i>	Predicted	<i>Corynebacterium glutamicum</i> ATCC 13032	<i>cg2888</i>	<i>cg1568</i>

The example in Table 17 presents a regulation being predicted twice by two different regulations of the same organism, *Corynebacterium glutamicum* ATCC 13032. In this

example, both genes of the target organism, *Corynebacterium efficiens* YS-314, are paralogs, which explains why the same regulation is predicted twice with the same source genes and organism.

Other modifications were made to the transfer pipeline, such as how the transfer pipeline deals with regulations of transcription units. Sometimes, as already discussed on the TAXI section, the transfer of transcription units among organisms can generate differentiations between the units in different organisms.

This differentiation can modify the position of the orthologue gene in the transcription unit of the organism that is receiving the regulation; the gene can become the second gene of the transcription unit, not being able to receive the binding motif that attracts the transcription factor, making the regulation impossible. For this special case, the new pipeline deals with a main regulation and a side regulation; the main regulation is the orthologue gene, which is participating in the process of transferring a regulation, and the first gene of the transcription unit is the side regulation because the upstream of this gene might make the existence of the binding motif to control the regulation possible.

There are three types of transfer of regulations that cover all possible cases with which the transfer pipeline has to deal: the normal transfer in which the target gene in the organism that receives the regulation is a monocistronic transcription unit. In other words, the target gene is transcribed alone. The second type deals with a normal transfer of a regulation, in which the target gene of the organism that is receiving the regulation is the first gene of the transcription unit, making the search for the binding motif and the completion of the transfer of regulation possible.

The third type is the trickiest one, since it deals with the main and side regulations; this type is subdivided into three types, depending of where the binding motif is found. In the first subtype, the binding motif is found ahead of both target genes, the target gene on the organism, which is receiving the regulation and the first gene of the transcription unit. In the second type, a predicted binding motif is found ahead only of the main regulation, or in other words, only in the upstream sequence of the orthologue gene on the target organism is found. The last type is the contrary of the second one, when the binding motif is found only on the upstream sequence of the first gene of the transcription unit. The following figures present examples of regulations transfer.

Figure 71 presents the normal transfer of regulation from a source organism to a target one. In this example case, a regulation was transferred from the organism

Corynebacterium glutamicum ATCC 13032 to the bacteria *Corynebacterium efficiens* YS-314, using the activation of transcription of the gene *cg0085* by the transcription factor *cg2888* as a source regulation, a validated regulation with two binding motifs and published in [105].

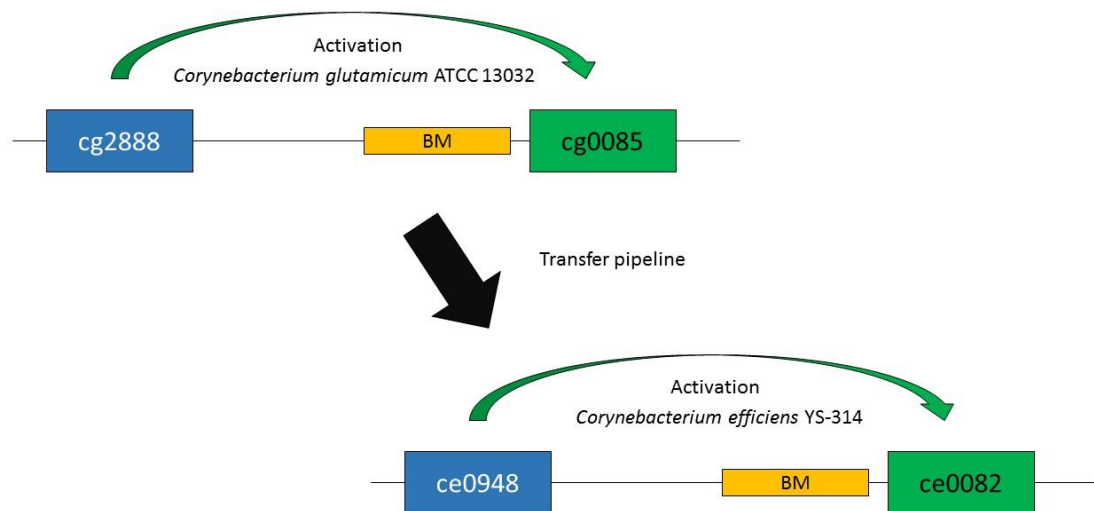


Figure 71: In the figure, the transfer of a regulation from the source organism *Corynebacterium glutamicum* ATCC 13032 to the target organism *Corynebacterium efficiens* YS-314 is seen. This transfer fits in the first type of transfer, in which the target gene of the organism that receives the regulation is a monocistronic transcription unit.

In the example presented in Figure 71, a regulation from the organism *Corynebacterium glutamicum* ATCC 13032 was transferred to the organism *Corynebacterium efficiens* YS-314; this transfer fits in the first type of transfer, since the target gene on the recipient organism is a monocistronic transcription unit.

The second type of transfer, as discussed already, covers the transfer of regulations to a target gene that is inserted on a polycistronic transcription unit, in which the target gene is also the first gene of the transcription unit. Figure 72 shows an example of this type of transference.

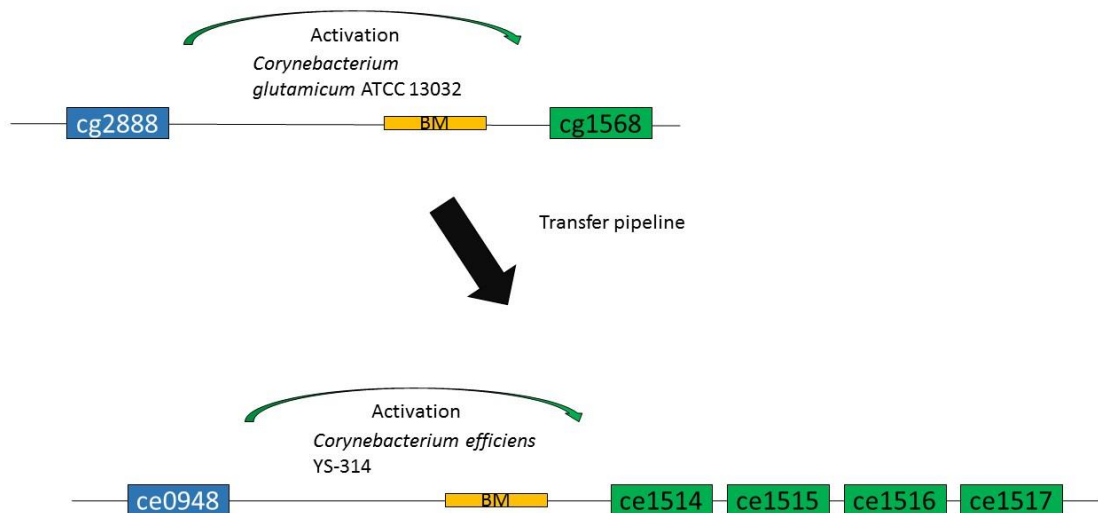


Figure 72: This figure represents an example of the second type of transfer covered by the new transfer pipeline, in which the target gene of the recipient organism is the first gene of a polycistronic transcription unit.

Again, in this example, a regulation from *Corynebacterium glutamicum* ATCC 13032 was transferred to the target organism *Corynebacterium efficiens* YS-314. The regulation of the monocistronic transcription unit of the gene *cg1568* [105] was transferred to the polycistronic transcription unit *op_ce1514*, predicting the regulation of the transcription unit by the predicted transcription factor *ce0948*. The regulation of the first gene of the transcription unit is stored as the main regulation of the TU, while all other genes, which are part of the transcription unit, also have their regulations stored but indicate the regulation of the first gene of the transcription unit.

The last type is represented in Figure 73; this image presents a regulation from *Corynebacterium glutamicum* ATCC 13032 being transferred to the organism *Corynebacterium efficiens* YS-314. The source regulation uses the same transcription factor as the previous examples but with a different target gene, as published in [105].

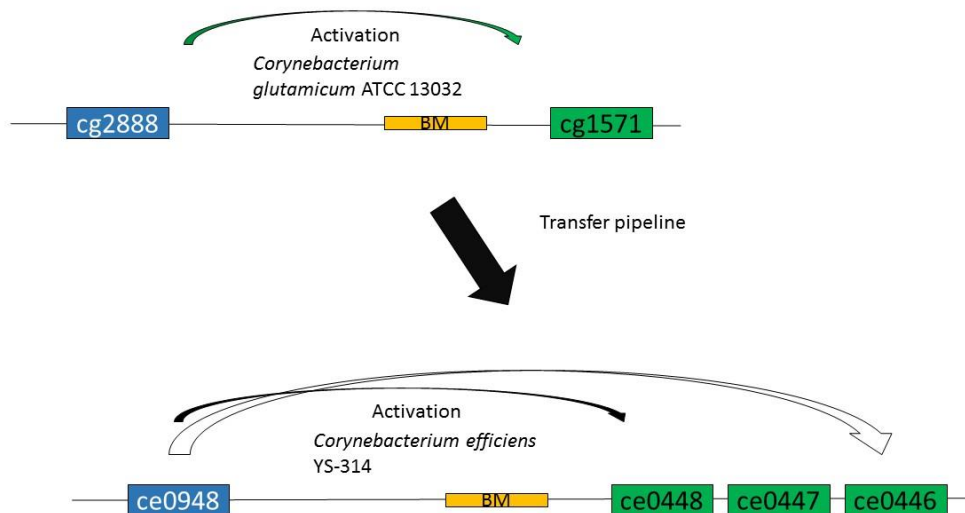


Figure 73: The transfer of a regulation from the organism *Corynebacterium glutamicum* ATCC 13032 to the organism *Corynebacterium efficiens* YS-314. The figure shows two arrows that indicate the presence of a main regulation and a side regulation, in which the orthologue target gene of the original organism is not the first gene of the transcription unit.

Figure 73 displays a regulation being transferred from *Corynebacterium glutamicum* ATCC 13032 to the bacterial organism *Corynebacterium efficiens* YS-314, originally on the source organism. This regulation is a transcription factor regulating a monocistronic transcription unit. However, on the target organism, the orthologue target gene is part of a polycistronic transcription unit. In this special case, the orthologue target gene is not the first gene of this polycistronic transcription unit; so the pipeline deals with a search on the upstream sequence of both genes, on the upstream sequence of the orthologue target gene and on the upstream sequence of the first gene of the transcription unit.

In Figure 73, the white arrow indicates the transfer of the regulation for the first gene of the transcription unit; this regulation was stored in the database with a flag indicating that this transferred regulation is a side regulation. In other words, the target gene actually is not the orthologue gene. The orthologue gene is also stored with a flag indicating that this is the main regulation. All other genes of the transcription unit have their regulations stored, but in this case, not only the regulation of the first gene of transcription unit is indicated but the main and side regulations are also indicated.

After running the pipeline with all the new features, a total amount of 3,971 regulations were achieved, 1/3 more predictions than the previous CoryneRegNet's version; the database also has a total of 4,993 evidences for predicted regulations. For this

CoryneRegNet’s version, only regulations of transcription factors were considered for transferring, excluding all regulations controlled by sigma factors from the pipeline. Table 18 presents the total of predicted regulations reached by CoryneRegNet’s back-end pipeline.

Table 18: The relation of organisms with the respective amounts of genes, predicted transcription factors, predicted target genes, predicted regulations, and predicted binding motifs.

Organism	Genes	TFs	Target genes	Regulations	BMs
<i>Corynebacterium jeikeium</i> K411	2104	22	148	191	151
<i>Corynebacterium glutamicum</i> R	3052	69	519	873	761
<i>Corynebacterium pseudotuberculosis</i> 1002	2057	31	249	375	245
<i>Corynebacterium urealyticum</i> DSM 7109	2024	19	157	221	157
<i>Corynebacterium kroppenstedtii</i> DSM 44385	2018	19	170	243	177
<i>Corynebacterium diphtheriae</i> NCTC 13129	2272	31	163	222	207
<i>Corynebacterium glutamicum</i> ATCC 13032	3057	6	86	87	38
<i>Escherichia coli</i> K12 MG1655	4237	-	-	-	-
<i>Corynebacterium pseudotuberculosis</i> C231	2053	31	249	365	245
<i>Corynebacterium aurimucosum</i> ATCC 700975	2531	29	227	349	265
<i>Corynebacterium efficiens</i> YS-314	2950	44	450	671	443
<i>Corynebacterium pseudotuberculosis</i> FRC41	2110	31	254	374	254
Total	30466	332	2672	3971	2943

3.4.3. Novel regulatory networks

New biologically-validated data and a new transfer pipelines generated novel regulatory networks for organisms stored in the CoryneRegNet database. Since the new prediction pipeline transfers regulations from any source organism to any target organism, only *Escherichia coli* K-12 MG1655 did not received regulations.

This section will address the gene regulatory network of three organisms that presents the biological nuances of *Corynebacterium glutamicum* ATCC 13032, *Escherichia coli* K12 MG1655, and *Corynebacterium pseudotuberculosis* FRC41. For the two first organisms, the updated data of the validated regulatory networks were used, and the third organism only presented a transferred regulatory network using all other validated data to predict the network as source.

3.4.3.1. *Corynebacterium glutamicum* ATCC 13032

For CoryneRegNet's first version, the bacterial organism *Corynebacterium glutamicum* ATCC 13032 was used as a model organism to study the composition of the bacteria with the regulatory networks generated by the transcription factors and target genes.

In CoryneRegNet's versions, the regulatory network was updated with modifications that added new regulations and, until version 6.0 of the system, just one organism was used as the source organism for the transfer pipeline. Table 19 presents the quantity of regulations in CoryneRegNet's 6.0 version for *Corynebacterium glutamicum* ATCC 13032.

Table 19: Presents the quantity of regulations for the organism *Corynebacterium glutamicum* ATCC 13032 in both levels of the database, the experimental data, a biologically validated data, and a second level with predicted information of CoryneRegNet’s back-end pipeline.

Version	Genes	TFs	Target genes	Regulations	BMs
Experimental	3058	98	786	1441	528
Predicted	3058	-	-	-	-

Table 19 presents the regulations for the organism *Corynebacterium glutamicum* ATCC 13032; there are no predicted regulations, since there is no transfer of regulations from other organisms targeting this organism. Table 20 presents the regulations present in the same organism: an updated version of biologically-validated data with the result of new transfer pipeline with regulations transferred from other organisms to this specific organism.

Table 20: Presents the quantity of regulations for the organism *Corynebacterium glutamicum* ATCC 13032 in both levels of the database, the validated data, and a second level with predicted information from CoryneRegNet’s back-end pipeline.

Version	Genes	TFs	Target genes	Regulations	BMs
Validated	3058	98	2336	2800	2645
Predicted	3058	6	86	87	38

After replacing the old transfer pipeline with the new one, some transfers appeared from other organisms to *Corynebacterium glutamicum* ATCC 13032, since the regulations from all other organisms are used in the transfer. In Table 20, the presence of 87 new regulations that came from other organisms can be seen.

Table 21 lists some examples of transferred regulations with the source organisms and the evidences that confirm the prediction; as discussed before, a prediction could be predicted more than once and then the evidence was stored.

Table 21: Regulations transferred from other organisms to *Corynebacterium glutamicum* ATCC 13032.

Source gene	Target gene	Source organism	Source gene (SO)	Target gene (SO)	Qty. of Evidence
<i>cg2103</i>	<i>cg3303</i>	<i>Corynebacterium efficiens</i> YS-314	<i>ce1812</i>	<i>ce2815</i>	1
<i>cg2103</i>	<i>cg0569</i>	<i>Corynebacterium efficiens</i> YS-314	<i>ce1812</i>	<i>ce1940</i>	1
<i>cg2103</i>	<i>cg0464</i>	<i>Corynebacterium efficiens</i> YS-314	<i>ce1812</i>	<i>ce1940</i>	1
<i>cg2103</i>	<i>cg2445</i>	<i>Corynebacterium diphtheriae</i> NCTC 13129	<i>dip1414</i>	<i>dip1669</i>	2
<i>cg0001</i>	<i>cg0699</i>	<i>Escherichia Coli</i> K12 MG1655	<i>b3702</i>	<i>b2508</i>	1

Table 21 presents five examples of transferred regulations; among these, there are some different types of transferring, covering some of the types of transfer already discussed.

Except for the forth example, all others are assorted as the most common type of transferred regulation, in which a regulation is directly transferred from one source organism to the target organism, and the target gene of the target organisms are not inside the transcription units.

The forth example is classified as the trickiest transfer type. The source regulations points to a gene that is not the first gene of the transcription unit. In other words, the main regulation is not actually the regulation of the transcription unit, existing aside the regulation for the first gene of the transcription unit.

With the evidence of the transcription of example four, there is a biologically-validated evidence for these regulations, characterizing the prediction as a confirmation of the validated evidence. Figure 74 presents this regulation, exemplifying the case of regulation transfer.

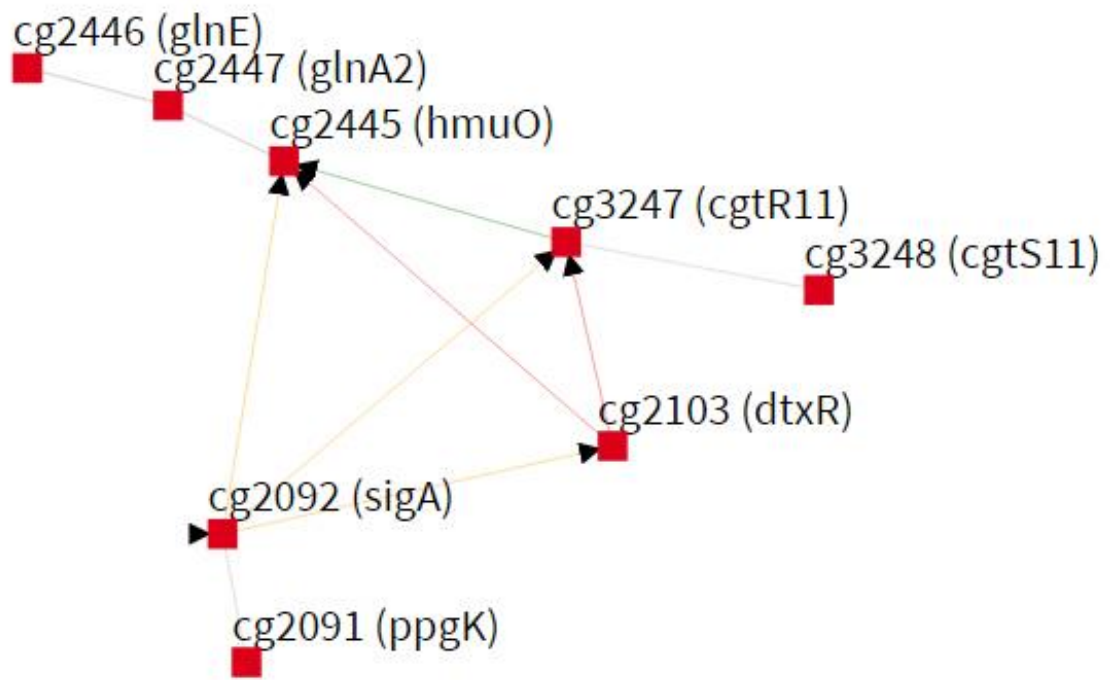


Figure 74: Figure representing the regulation of the gene *cg2445*. It presents each regulation related to the gene, not only of transcription factors, but also the regulation by sigma factors.

Figure 74 presents the regulation of the gene “*cg2445*”. It presents the regulation of transcription factors and sigma factors with the interconnection inside the same transcription unit. The regulation of repression of transcription of the gene “*cg2445*” by the gene “*cg2103*” is already validated [106], comprising a prediction confirming a validated regulation. Figure 75 shows the complete gene regulatory network of the bacterial genome of *Corynebacterium glutamicum* ATCC 13032.

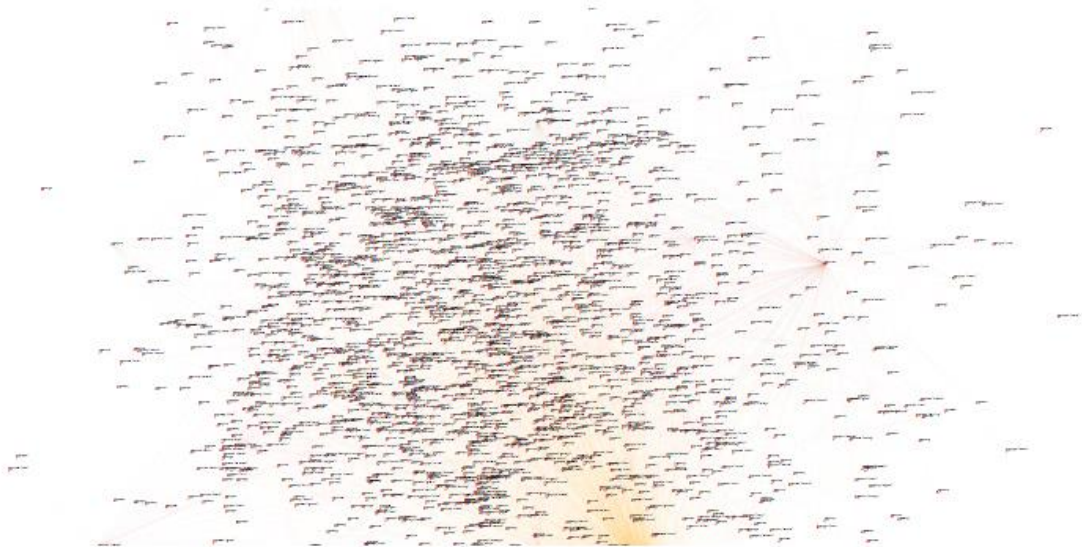


Figure 75: Figure presenting the entire regulatory network for the organism *Corynebacterium glutamicum* ATCC 13032, with 2,800 regulations. The complete regulatory network involves almost all genes of the bacterial genome.

Figure 75 presents the entire gene regulatory network for the organism *Corynebacterium glutamicum* ATCC 13032. It presents the regulations controlled by the transcription factors and sigma factors following the scheme previously discussed. The graph size is determined by the amount of genes inserted in the graph, and the distance between the genes varies if the link is a representation of a connection of a regulation or connection between genes inside the same transcription unit.

3.4.3.2. *Escherichia coli* K12 MG1655

A model organism that was inserted later was the proteobacteria *Escherichia coli* K12 MG1655; this organism is also a widely well-studied organism with a major project that analyzes every aspect of the bacterial genome. RegulonDB [24], an already discussed database, shows information regarding transcription units and regulatory networks alongside other features of the bacteria.

RegulonDB's current version, 8.0 [24], presents an improvement of studies on the regulatory network for this bacteria. Table 22 presents the amount of regulations used on CoryneRegNet's previous version.

Table 22: Regulations for *Escherichia coli* K12 MG1655 present in CoryneRegNet's 6.0 version.

Version	Genes	TFs	Target genes	Regulations	BMs
Experimental	4237	144	1102	2245	1219
Predicted	4237	-	-	-	-

Moreover, as observed for *Corynebacterium glutamicum* ATCC 13032, there are no predicted regulations resulted from the transfer pipeline from other organisms to *Escherichia coli* K12 MG1655. Next, Table 23 presents the amount of regulations used in this CoryneRegNet's update; all regulations were extracted from RegulonDB's current version. Also, on the same table, the results of CoryneRegNet's back-end run for this organism were presented with a further explanation.

Table 23: Updated amount of regulations for *Escherichia coli* K12 MG1655 extracted from RegulonDB with the result of CoryneRegNet's backend run for this organism.

Version	Genes	TFs	Target genes	Regulations	BMs
Validated	4237	206	3486	6413	4917
Predicted	4237	-	-	-	-

A particularity presented by the run of CoryneRegNet's back-end for this organism was the absence of predicted regulations. This fact can be explained by the phylogenetic distance between the proteobacteria (*Escherichia coli* K12 MG1655) *phylum* and Actinobacteria (*Corynebacterium*). This distance generated a lower quantity of shared clusters that influences directly the transfer pipeline. Figure 76 presents the full-gene regulatory network for this organism.

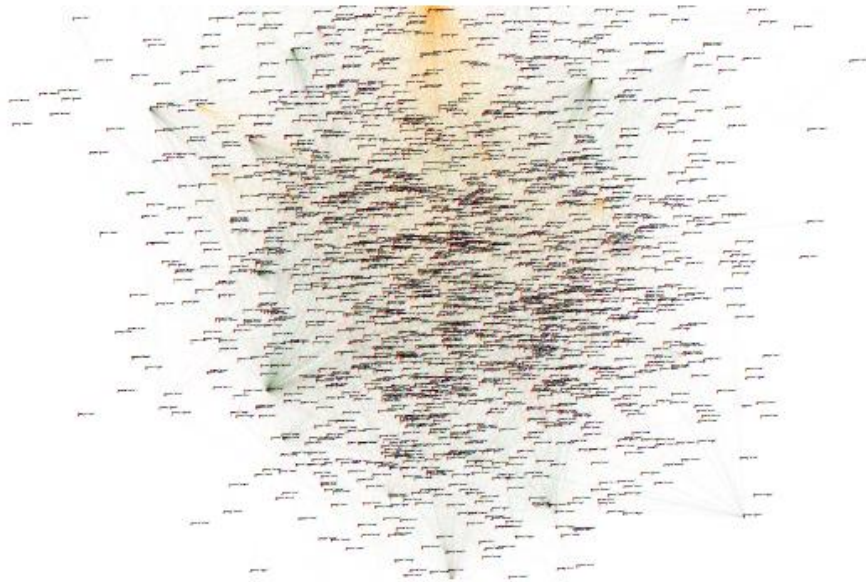


Figure 76: Figure presenting the full-gene regulatory network for the bacterial organism *Escherichia coli* K12 MG1655, with regulations controlled by transcription factors and sigma factors.

3.4.3.3. *Corynebacterium pseudotuberculosis* FRC41

Inserted in CoryneRegNet’s current version, this organism does not have biologically-validated information of gene regulatory networks, presenting only predicted results of CoryneRegNet’s back-end run. Table 24 presents the predicted information of the gene regulatory network for this organism in CoryneRegNet’s current version.

Table 24: Regulations for *Corynebacterium pseudotuberculosis* FRC41 present in CoryneRegNet’s 6.0 version.

Version	Genes	TFs	Target genes	Regulations	BMs
Experimental	2110	-	-	-	-
Predicted	2110	30	214	291	175

Limited by only transferring regulations from *Corynebacterium glutamicum* ATCC 13032, all predicted regulations of the *Corynebacterium pseudotuberculosis* FRC41 bacterial organism are comes from this organism. CoryneRegNet’s update opened the possibility of predicting new regulations for this organism using a larger amount of organisms as source. Table 25 presents the result of CoryneRegNet’s back-end run.

Table 25: Table showing the result of CoryneRegNet’s back-end run predicting regulations for the organism *Corynebacterium pseudotuberculosis* FRC41.

Version	Genes	TFs	Target genes	Regulations	BMs
Validated	2110	-	-	-	-
Predicted	2110	31	254	374	254

Replacing the transfer pipeline with the new one resulted in the easily-noticed increase of 30% on predictions for this organism, creating a whole new predicted regulatory network with 374 regulations. Table 26 presents some examples of transferred regulations from other organisms to *Corynebacterium pseudotuberculosis* FRC41.

Table 26: Table presenting examples of predicted regulations for the organism *Corynebacterium pseudotuberculosis* FRC41 from different source organisms.

Source gene	Target gene	Source organism	Source gene (SO)	Target gene (SO)	Qty. of Evidence
<i>cpfr_1</i>	<i>cpfr_1893</i>	<i>Escherichia coli</i> K12 MG1655	<i>b3702</i>	<i>b1415</i>	1
<i>cpfr_205</i>	<i>cpfr_1429</i>	<i>Corynebacterium glutamicum</i> ATCC 13032	<i>cg0350</i>	<i>cg2410</i>	1
<i>cpfr_1645</i>	<i>cpfr_1647</i>	<i>Corynebacterium glutamicum</i> ATCC 13032	<i>cg2737</i>	<i>cg0957</i>	1
<i>cpfr_205</i>	<i>cpfr_891</i>	<i>Corynebacterium glutamicum</i> ATCC 13032	<i>cg0350</i>	<i>cg1435</i>	1
<i>cpfr_1525</i>	<i>cpfr_1290</i>	<i>Corynebacterium glutamicum</i> ATCC 13032	<i>cg2502</i>	<i>cg0794</i>	1

The five examples in Table 26 are a little representation of all predicted regulations, which were transferred from other organisms to *Corynebacterium pseudotuberculosis* FRC41. The example of predicted regulation of the predicted transcription factor “*cpfr_1525*” repressing gene “*cpfr_1290*”, transferred from the organism *Corynebacterium glutamicum* ATCC 13032, as the example discussed, is an example of a regulation with main and side regulations for *Corynebacterium glutamicum* ATCC

13032. Figure 77 has a representation of the regulation alongside all regulations around this specific one.

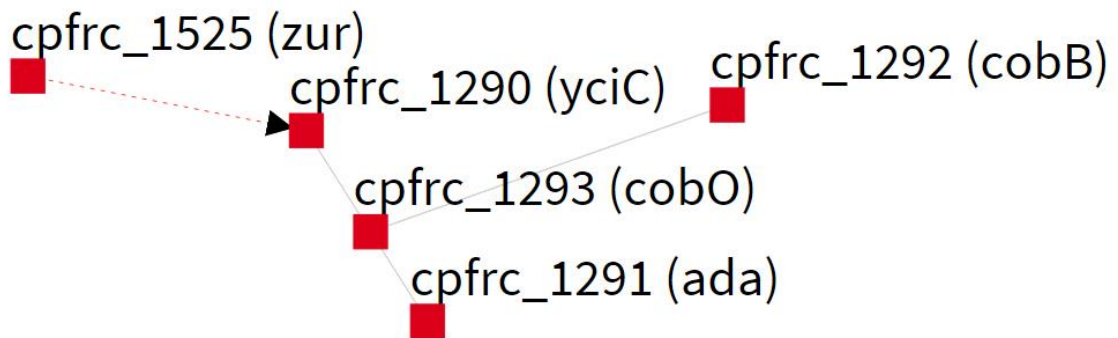


Figure 77: Figure showing the predicted regulation of the predicted transcription factor cpfrc_1525 repressing the transcription of cpfrc_1290.

This regulation represented is a transferred regulation, targeting a gene inside the transcription unit, not the first gene, characterizing a transference comprising a main regulation and a side regulation. However, in this special case, the transfer pipeline did not predict a binding motif ahead of the gene on the upstream sequence of the first gene on the transcription unit. The transfer was validated because the main regulation exists, a regulation that is shown on Figure 77.

3.5. Conclusion

The fast growth of biological studies generates a big amount of data, which have to be compiled and stored in a way that facilitates the access and comprehension of the data by the researcher.

CoryneRegNet's system is not immune to the problems generated by the data growth. With new studies on gene regulatory networks for model organisms, the acquisition of this data by the system increased the response time of the consults performed on the back-end database. This problem resulted in the demand that the system might be updated, so that it might present an acceptable performance again.

The update of the system went by different steps, whose main goal was to maintain the same reliability of previous CoryneRegNet's versions. The process resulted in the complete redesign of the database, specifically developed to satisfy the necessities of the system. A totally new back-end program assisted the creation of the database, eliminating the interference of human errors during its creation and replacing old tools for new and modern ones, keeping CoryneRegNet on the vanguard and finally generating a totally new web-based, user-friendlier front-end.

CoryneRegNet's new version presents a new look to handle the same data, with the same quality as the previous versions, but additionally creating the foundations for a brighter future for the system with new acquired data.

4. CONCLUSION

The study of bacterial and archaeal organisms supplements the foundation to understand the basis of life. These primary organisms presented, during their evolution, adaptations that were preformed to sustain life.

Surviving in different environments or with different organisms in same spaces forced the evolution of these organisms, sharing and creating new mechanisms. With the newcomers, the organism generated control agents to regulate the new mechanisms.

The basal control of the transcription of these mechanisms was linked to the already-existing control of other mechanisms of the bacteria or archaea, becoming a part of an already-existing gene regulatory network.

Both studies complement themselves; the understanding of a gene acquisition process by an organism could explain the creation of a gene regulatory network in which an organism is inserted in a new environment and it has to respond to external biological pressures.

5. OUTLOOK

The main goal of the project was to generate resources to researchers to perform studies through genomes of *bacteria* and *archaea*. By studying the evolution and the regulatory networks of these organisms, the researcher can understand how they survive and grow with the use of less expensive methods.

The development of TAXI gives a systems biology tool to the user to compare the evolution of related and non-related organisms, to understand the acquisition of genes through the taxonomy evolution of the organisms or speciation.

The tool provides, for the user, an easy access to statistical information, besides graphics and queries that may be performed on the organisms, helping the user to understand the acquisition of genes made by the organism and the structure of transcription units, elucidating genes and transcription units that were exclusive to a specific species or strain.

On the other hand, the update of an already well-known database, CoryneRegNet, raise this systems biology tool to a new level of development; new tools and features bring a higher level of accuracy to the transferring regulatory networks, from a donor organism to a receptor organism.

The update process went through the development of a new back-end, based on CoryneRegNet's old version, following the same steps but with more accurate tools to perform the same tasks.

Also, a new back-end database was developed to sustain the development of the system and to open a branch to add new organisms to the database, without decreasing the performance and maintaining a good level of speed for the front-end.

The last step of CoryneRegNet's update was the development of a new front-end for the system, also with new features and tools combined with a new, user-friendlier interface in which the user could perform researches.

6. FUTURE WORK

Both tools created and updated on this work, fomented the bases for studies on speciation of bacteria and archaea, alongside the study of gene regulatory networks among related organisms.

The development of both tools it is not completed, lasting some features which still needing to be implemented to help the user during the research on biological information hosted on both databases.

For TAXI;

- Implementation of graphs to present graphically the speciation process of a transcription unit.
- Implementation of comparisons among different types of gene clusters, UEKO (less restrict) x MOG (more restrict).
- Calculation of orthologue transcription units, use the already created database with operon prediction and gene orthologue groups, to generate groups of orthologue transcription units.

For CoryneRegNet;

- Implementation of genome browser, use the already created database with genome annotation, alongside a library in JavaScript, or similar code language, to generate a navigation browser among the organism.

For Both:

- Perform a study over the evolution of regulatory networks among the speciation, identifying regulations possibly created or modified with the speciation process.

7. BIBLIOGRAPHY

- [1] H. Salgado, G. Moreno-Hagelsieb, T. F. Smith, and J. Collado-Vides, "Operons in Escherichia coli: genomic analyses and predictions.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 12, pp. 6652–7, 2000.
- [2] J. G. Lawrence and J. R. Roth, "Selfish operons: Horizontal transfer may drive the evolution of gene clusters," *Genetics*, vol. 143, no. 4, pp. 1843–1860, 1996.
- [3] J. G. Lawrence, "Selfish operons and speciation by gene transfer.," *Trends Microbiol.*, vol. 5, no. 9, pp. 355–9, Sep. 1997.
- [4] W. Davids and Z. Zhang, "The impact of horizontal gene transfer in shaping operons and protein interaction networks – direct evidence of preferential attachment," *BMC Evol. Biol.*, vol. 8, no. 1, p. 23, 2008.
- [5] G. Moreno-Hagelsieb, V. Treviño, E. Pérez-Rueda, T. F. Smith, and J. Collado-Vides, "Transcription unit conservation in the three domains of life: a perspective from Escherichia coli.," *Trends Genet.*, vol. 17, no. 4, pp. 175–7, Apr. 2001.
- [6] T. Opperman and J. P. Richardson, "Phylogenetic analysis of sequences from diverse bacteria with homology to the Escherichia coli rho gene.," *J. Bacteriol.*, vol. 176, no. 16, pp. 5033–43, 1994.
- [7] I. Erill, M. Jara, N. Salvador, M. Escribano, S. Campoy, and J. Barbé, "Differences in LexA regulon structure among Proteobacteria through in vivo assisted comparative genomics," *Nucleic Acids Res.*, vol. 32, no. 22, pp. 6617–6626, 2004.
- [8] C. a. Ouzounis and P. D. Karp, "Global properties of the metabolic map of Escherichia coli," *Genome Res.*, vol. 10, no. 4, pp. 568–576, 2000.
- [9] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of Escherichia coli.," *Nat. Genet.*, vol. 31, no. 1, pp. 64–8, 2002.
- [10] D. J. Studholme, M. Buck, and B. T. Nixon, "Identification of potential ??N -dependent promoters in bacterial genomes," *Microbiology*, vol. 146, no. 12, pp. 3021–3023, 2000.
- [11] I. Cases, D. W. Ussery, and V. De Lorenzo, "The σ_{54} regulon (sigmulon) of Pseudomonas putida," *Environ. Microbiol.*, vol. 5, no. 12, pp. 1281–1293, 2003.
- [12] T. S. Rani and R. S. Bapi, "Analysis of n-gram based promoter recognition methods and application to whole genome promoter prediction.," *In Silico Biol.*, vol. 9, no. 1–2, pp. S1–16, 2009.
- [13] S. Mann, J. Li, and Y.-P. P. Chen, "A pHMM-ANN based discriminative approach to promoter identification in prokaryote genomic contexts," *Nucleic Acids Res.*, vol. 35, no. 2, pp. e12–e12, Dec. 2006.
- [14] S. Burden, Y.-X. Lin, and R. Zhang, "Improving promoter prediction for the NNPP2.2 algorithm: a case study using Escherichia coli DNA sequences.," *Bioinformatics*, vol. 21, no. 5, pp. 601–7, Mar. 2005.
- [15] C. Bland, A. S. Newsome, and A. A. Markovets, "Promoter prediction in E. coli based on SIDD profiles and Artificial Neural Networks," *BMC Bioinformatics*, vol. 11, no. Suppl 6, p. S17, 2010.
- [16] S. de Avila E Silva, S. Echeverrigaray, and G. J. L. Gerhardt, "BacPP: bacterial promoter prediction--a tool for accurate sigma-factor specific assignment in enterobacteria.," *J. Theor. Biol.*, vol. 287, pp. 92–9, Oct. 2011.
- [17] M. W. Towsey, J. J. Gordon, and J. M. Hogan, "The prediction of bacterial transcription start sites using SVMs.," *Int. J. Neural Syst.*, vol. 16, no. 5, pp. 363–70, Oct. 2006.
- [18] M. Towsey, P. Timms, J. Hogan, and S. A. Mathews, "The cross-species prediction of bacterial promoters using a support vector machine.," *Comput. Biol. Chem.*, vol. 32, no. 5, pp. 359–66, Oct. 2008.
- [19] J. J. Gordon, M. W. Towsey, J. M. Hogan, S. A. Mathews, and P. Timms, "Improved prediction of bacterial transcription start sites," *Bioinformatics*, vol. 22, no. 2, pp. 142–148, 2006.
- [20] D. a Fell and a Wagner, "The small world of metabolism.," *Nat. Biotechnol.*, vol. 18, no. 11, pp. 1121–1122, 2000.
- [21] M. J. Herrgård, M. W. Covert, and B. Ø. Palsson, "Reconstruction of microbial transcriptional regulatory networks," *Curr. Opin. Biotechnol.*, vol. 15, no. 1, pp. 70–77, 2004.
- [22] A. Martínez-Antonio and J. Collado-Vides, "Identifying global regulators in transcriptional regulatory

- networks in bacteria,” *Curr. Opin. Microbiol.*, vol. 6, no. 5, pp. 482–489, 2003.
- [23] I. Lozada-Chavez, “Bacterial regulatory networks are extremely flexible in evolution,” *Nucleic Acids Res.*, vol. 34, no. 12, pp. 3434–3445, 2006.
- [24] H. Salgado, M. Peralta-Gil, S. Gama-Castro, A. Santos-Zavaleta, L. Muñiz-Rascado, J. S. García-Sotelo, V. Weiss, H. Solano-Lira, I. Martínez-Flores, A. Medina-Rivera, G. Salgado-Osorio, S. Alquicira-Hernández, K. Alquicira-Hernández, A. López-Fuentes, L. Porrón-Sotelo, A. M. Huerta, C. Bonavides-Martínez, Y. I. Balderas-Martínez, L. Pannier, M. Olvera, A. Labastida, V. Jiménez-Jacinto, L. Vega-Alvarado, V. Del Moral-Chávez, A. Hernández-Alvarez, E. Morett, and J. Collado-Vides, “RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more.,” *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D203–13, 2013.
- [25] N. Sierro, Y. Makita, M. de Hoon, and K. Nakai, “DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information.,” *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D93–6, 2008.
- [26] E. J. Alm, K. H. Huang, M. N. Price, R. P. Koche, K. Keller, I. L. Dubchak, and A. P. Arkin, “The MicrobesOnline Web site for comparative genomics,” *Genome Res.*, vol. 15, no. 7, pp. 1015–1022, 2005.
- [27] A. Grote, J. Klein, I. Retter, I. Haddad, S. Behling, B. Bunk, I. Biegler, S. Yarmolinetz, D. Jahn, and R. Münch, “PRODORIC (release 2009): A database and tool platform for the analysis of gene regulation in prokaryotes,” *Nucleic Acids Res.*, vol. 37, no. October 2008, pp. 61–65, 2009.
- [28] D. L. Nelson and M. M. Cox, *Principles of biochemistry*, vol. 1. 2010.
- [29] F. Crick, “Central Dogma Of Molecular Biology,” *Nature*, vol. 227, pp. 561–563, 1970.
- [30] J. Tamames, G. Casari, C. Ouzounis, and A. Valencia, “Conserved Clusters of Functionally Related Genes in Two Bacterial Genomes,” *J. Mol. Evol.*, vol. 44, no. 1, pp. 66–73, Jan. 1997.
- [31] R. Overbeek, M. Fonstein, M. D’Souza, G. D. Pusch, and N. Maltsev, “The use of gene clusters to infer functional coupling.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 6, pp. 2896–2901, 1999.
- [32] Y. Zheng, J. D. Szustakowski, L. Fortnow, R. J. Roberts, and S. Kasif, “Computational Identification of Operons in Microbial Genomes Computational Identification of Operons in Microbial Genomes,” *Genome Res.*, pp. 1221–1230, 2002.
- [33] S. Ballouz, A. R. Francis, R. Lan, and M. M. Tanaka, “Conditions for the evolution of gene clusters in bacterial genomes.,” *PLoS Comput. Biol.*, vol. 6, no. 2, p. e1000672, 2010.
- [34] R. Fani, M. Brillì, and P. Liò, “The origin and evolution of operons: The piecewise building of the proteobacterial histidine operon,” *J. Mol. Evol.*, vol. 60, no. 3, pp. 378–390, 2005.
- [35] J. Lawrence, “Selfish operons: The evolutionary impact of gene clustering in prokaryotes and eukaryotes,” *Curr. Opin. Genet. Dev.*, vol. 9, no. 6, pp. 642–648, 1999.
- [36] C. . Orengo, T. P. Flores, W. R. Taylor, and J. M. Thornton, “Identification and classification of protein fold families,” *Protein Eng. Des. Sel.*, vol. 6, no. 5, pp. 485–500, 1993.
- [37] N. E. Stahl, Franklin, W., Murray, “The evolution of gene clusters and genetic circularity in microorganisms.,” *Genetics*, vol. 53, no. March, pp. 569–576, 1966.
- [38] R. Losick and L. Shapiro, “Changing views on the nature of the bacterial cell: From biochemistry to cytology,” *J. Bacteriol.*, vol. 181, no. 14, pp. 4143–4145, 1999.
- [39] T. Itoh, K. Takemoto, H. Mori, and T. Gojobori, “Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes.,” *Mol. Biol. Evol.*, vol. 16, no. 3, pp. 332–346, 1999.
- [40] B. Fran, D. Perrin, C. Sanchez, and J. Monod, “The Operon : A Group of Genes Whose Expression is Coordinated by an Operator < I loo loo,” *J. Bacteriol.*, vol. 1729, pp. 1727–1729, 1960.
- [41] G. Moreno-Hagelsieb and J. Collado-Vides, “A powerful non-homology method for the prediction of operons in prokaryotes.,” *Bioinformatics*, vol. 18 Suppl 1, pp. S329–S336, 2002.
- [42] B. Taboada, C. Verde, and E. Merino, “High accuracy operon prediction method based on STRING database scores,” *Nucleic Acids Res.*, vol. 38, no. 12, p. e130, 2010.
- [43] S. A. Kumar, “The structure and mechanism of action of bacterial DNA-dependent RNA polymerase.,” *Prog. Biophys. Mol. Biol.*, vol. 38, no. 3, pp. 165–210, 1981.
- [44] M. M. S. M. Wosten, “Eubacterial sigma-factors,” *FEMS Microbiol. Rev.*, vol. 22, pp. 127–150, 1998.
- [45] P. H. von Hippel, D. G. Bear, W. D. Morgan, and J. A. McSwiggen, “Protein-Nucleic Acid Interactions in

- Transcription: A Molecular Analysis,” *Annu. Rev. Biochem.*, vol. 53, no. 1, pp. 389–446, Jun. 1984.
- [46] D. F. Browning and S. J. Busby, “The regulation of bacterial transcription initiation,” *Nat. Rev. Microbiol.*, vol. 2, no. 1, pp. 57–65, Jan. 2004.
- [47] S. C. Janga, H. Salgado, J. Collado-Vides, and A. Martínez-Antonio, “Internal Versus External Effector and Transcription Factor Gene Pairs Differ in Their Relative Chromosomal Position in *Escherichia coli*,” *J. Mol. Biol.*, vol. 368, no. 1, pp. 263–272, Apr. 2007.
- [48] A. S. N. Seshasayee, G. M. Fraser, M. Madan Babu, and N. M. Luscombe, “Principles of transcriptional regulation and evolution of the metabolic system in *E. coli*,” *Genome Res.*, vol. 19, no. 1, pp. 79–91, 2009.
- [49] M. M. Babu and S. A. Teichmann, “Evolution of transcription factors and the gene regulatory network in *Escherichia coli*,” *Nucleic Acids Res.*, vol. 31, no. 4, pp. 1234–1244, 2003.
- [50] V. Anantharaman, E. V. Koonin, and L. Aravind, “Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains,” Edited by F. Cohen,” *J. Mol. Biol.*, vol. 307, no. 5, pp. 1271–1292, Apr. 2001.
- [51] M. J. Merrick and R. A. Edwards, “Nitrogen control in bacteria,” *Microbiol. Rev.*, vol. 59, no. 4, pp. 604–622, 1995.
- [52] M. Madan Babu, S. A. Teichmann, and L. Aravind, “Evolutionary Dynamics of Prokaryotic Transcriptional Regulatory Networks,” *J. Mol. Biol.*, vol. 358, no. 2, pp. 614–633, Apr. 2006.
- [53] A. Ishihama, “Molecular assembly and functional modulation of *Escherichia coli* RNA polymerase,” *Adv. Biophys.*, vol. 26, pp. 19–31, 1990.
- [54] J. Pena-Sanchez, S. Poggio, U. Flores-Perez, A. Osorio, C. Domenzain, G. Dreyfus, and L. Camarena, “Identification of the binding site of the 54 hetero-oligomeric FleQ/FleT activator in the flagellar promoters of *Rhodobacter sphaeroides*,” *Microbiology*, vol. 155, no. 5, pp. 1669–1679, 2009.
- [55] R. A. Mooney, S. A. Darst, and R. Landick, “Sigma and RNA Polymerase: An On-Again, Off-Again Relationship?,” *Mol. Cell*, vol. 20, no. 3, pp. 335–345, 2005.
- [56] J. D. Helmann, “Compilation and analysis of *Bacillus Subtilis* σ A -dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA,” *Nucleic Acids Res.*, vol. 23, no. 13, pp. 2351–2360, 1995.
- [57] H. Barrios, B. Valderrama, and E. Morett, “Compilation and analysis of sigma(54)-dependent promoter sequences,” *Nucleic Acids Res.*, vol. 27, no. 22, pp. 4305–4313, 1999.
- [58] E. Potvin, F. Sanschagrin, and R. C. Levesque, “Sigma factors in *Pseudomonas aeruginosa*,” *FEMS Microbiol. Rev.*, vol. 32, no. 1, pp. 38–55, 2008.
- [59] C. B. Harley and R. P. Reynolds, “Analysis of *E. coli* promoter sequences,” *Nucleic Acids Res.*, vol. 15, no. 5, pp. 2343–2361, 1987.
- [60] L. T. Macneil and A. J. M. Walhout, “Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression,” pp. 645–657, 2011.
- [61] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, “Inferring Regulatory Networks from Expression Data Using Tree-Based Methods,” *PLoS One*, vol. 5, no. 9, p. e12776, 2010.
- [62] G. Karlebach and R. Shamir, “Modelling and analysis of gene regulatory networks,” *Nat. Rev. Mol. Cell Biol.*, vol. 9, no. 10, pp. 770–780, Oct. 2008.
- [63] J. S. Parkinson, “Signal transduction schemes of bacteria,” *Cell*, vol. 73, no. 5, pp. 857–871, 1993.
- [64] J. B. Stock, A. J. Ninfa, and M. Stock, *Protein phosphorylation and regulation of adaptive responses in bacteria.*, vol. 53, no. 4. 1989.
- [65] I. Matic, F. Taddei, and M. Radman, “Survival versus maintenance of genetic stability: A conflict of priorities during stress,” *Res. Microbiol.*, vol. 155, pp. 337–341, 2004.
- [66] M. T. Laub and M. Goulian, “Specificity in Two-Component Signal Transduction Pathways,” *Annu. Rev. Genet.*, vol. 41, no. 1, pp. 121–145, 2007.
- [67] A. M. Stock, V. L. Robinson, and P. N. Goudreau, “Two-component signal transduction,” *Annu. Rev. ...*, vol. 69, pp. 183–215, 2000.
- [68] P. S. Dehal, M. P. Joachimiak, M. N. Price, J. T. Bates, J. K. Baumohl, D. Chivian, G. D. Friedland, K. H. Huang, K. Keller, P. S. Novichkov, I. L. Dubchak, E. J. Alm, and A. P. Arkin, “MicrobesOnline: An integrated portal for comparative and functional genomics,” *Nucleic Acids Res.*, vol. 38, no. SUPPL.1, pp.

396–400, 2009.

- [69] M. Pertea, K. Ayanbule, M. Smedinghoff, and S. L. Salzberg, “OperonDB: a comprehensive database of predicted operons in microbial genomes.,” *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D479–82, 2009.
- [70] F. Mao, P. Dam, J. Chou, V. Olman, and Y. Xu, “DOOR: A database for prokaryotic operons,” *Nucleic Acids Res.*, vol. 37, no. SUPPL. 1, pp. 459–463, 2009.
- [71] S. Okuda, T. Katayama, S. Kawashima, S. Goto, and M. Kanehisa, “ODB: a database of operons accumulating known operons across multiple genomes.,” *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D358–D362, 2006.
- [72] S. Okuda and A. C. Yoshizawa, “ODB: A database for operon organizations, 2011 update,” *Nucleic Acids Res.*, vol. 39, no. SUPPL. 1, pp. 552–555, 2011.
- [73] B. Taboada, R. Ciria, C. E. Martinez-Guerrero, and E. Merino, “ProOpDB: Prokaryotic Operon DataBase.,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D627–31, 2012.
- [74] G. R. Fernandes, D. V. Barbosa, F. Prosdociami, I. a. Pena, L. Santana-Santos, O. Coelho Junior, a. Barbosa-Silva, H. M. Velloso, M. a. Mudado, D. a. Natale, a. C. Faria-Campos, S. C. Aguiar, and J. M. Ortega, “A procedure to recruit members to enlarge protein family databases--the building of UECOG (UniRef-Enriched COG Database) as a model.,” *Genet. Mol. Res.*, vol. 7, no. 3, pp. 910–924, 2008.
- [75] M. N. Price, K. H. Huang, E. J. Alm, and A. P. Arkin, “A novel method for accurate operon predictions in all sequenced prokaryotes,” *Nucleic Acids Res.*, vol. 33, no. 3, pp. 880–892, 2005.
- [76] R. Münch, K. Hiller, A. Grote, M. Scheer, J. Klein, M. Schobert, D. Jahn, and R. Mu, “Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes.,” *Bioinformatics*, vol. 21, no. 22, pp. 4187–9, 2005.
- [77] J. Klein, S. Leupold, R. Münch, C. Pommerenke, T. Johl, U. Kärst, L. Jänsch, D. Jahn, and I. Retter, “ProdoNet: identification and visualization of prokaryotic gene regulatory and metabolic networks.,” *Nucleic Acids Res.*, vol. 36, no. Web Server issue, pp. W460–4, 2008.
- [78] P.-E. Jacques, A. L. Gervais, M. Cantin, J.-F. Lucier, G. Dallaire, G. Drouin, L. Gaudreau, J. Goulet, and R. Brzezinski, “MtbRegList, a database dedicated to the analysis of transcriptional regulation in *Mycobacterium tuberculosis*.,” *Bioinformatics*, vol. 21, no. 10, pp. 2563–5, 2005.
- [79] J.-C. Camus, M. J. Pryor, C. Médigue, and S. T. Cole, “Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv.,” *Microbiology*, vol. 148, no. Pt 10, pp. 2967–73, 2002.
- [80] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. a. Natale, “The COG database: an updated version includes eukaryotes.,” *BMC Bioinformatics*, vol. 4, p. 41, 2003.
- [81] M. J. Cipriano, P. N. Novichkov, A. E. Kazakov, D. A. Rodionov, A. P. Arkin, M. S. Gelfand, and I. Dubchak, “RegTransBase - a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes,” *BMC Genomics*, vol. 14, p. 213, 2013.
- [82] P. S. Novichkov, D. A. Rodionov, E. D. Stavrovskaya, E. S. Novichkova, A. E. Kazakov, M. S. Gelfand, A. P. Arkin, A. A. Mironov, and I. Dubchak, “RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach,” *Nucleic Acids Res.*, vol. 38, no. Web Server, pp. W299–W307, 2010.
- [83] P. S. Novichkov, A. E. Kazakov, D. A. Ravcheev, S. A. Leyn, G. Y. Kovaleva, R. A. Sutormin, M. D. Kazanov, W. Riehl, A. P. Arkin, I. Dubchak, and D. A. Rodionov, “RegPrecise 3.0--a resource for genome-scale exploration of transcriptional regulation in bacteria.,” *BMC Genomics*, vol. 14, no. 1, p. 745, 2013.
- [84] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender, “TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes,” *Nucl. Acids Res.*, vol. 34, no. suppl_1, pp. D108–110, 2006.
- [85] O. V. Kel-Margoulis, A. E. Kel, I. Reuter, I. V. Deineko, and E. Wingender, “TRANSCompel: a database on composite regulatory elements in eukaryotic genes.,” *Nucleic Acids Res.*, vol. 30, no. 1, pp. 332–4, 2002.
- [86] a. E. Kel, E. Gößling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender, “MATCHTM: A tool for searching transcription factor binding sites in DNA sequences,” *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3576–3579, 2003.
- [87] D. S. Chekmenev, C. Haid, and a. E. Kel, “P-Match: Transcription factor binding site search by combining patterns and weight matrices,” *Nucleic Acids Res.*, vol. 33, no. SUPPL. 2, pp. 432–437, 2005.

- [88] A. de Jong, H. Pietersma, M. Cordes, O. P. Kuipers, and J. Kok, "PePPER: a webserver for prediction of prokaryote promoter elements and regulons," *BMC Genomics*, vol. 13, no. 1, p. 299, 2012.
- [89] A. D. González, V. Espinosa, A. T. Vasconcelos, E. Pérez-Rueda, and J. Collado-Vides, "TRACTOR_DB: A database of regulatory networks in gamma-proteobacterial genomes," *Nucleic Acids Res.*, vol. 33, no. DATABASE ISS., pp. 98–102, 2005.
- [90] A. G. Pérez, V. E. Angarica, A. T. R. Vasconcelos, and J. Collado-Vides, "Tractor_DB (version 2.0): A database of regulatory interactions in gamma-proteobacterial genomes," *Nucleic Acids Res.*, vol. 35, no. November 2006, pp. 132–136, 2007.
- [91] M. Pachkov, P. J. Balwierz, P. Arnold, E. Ozonov, and E. van Nimwegen, "SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates.," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D214–20, 2013.
- [92] P. Arnold, I. Erb, M. Pachkov, N. Molina, and E. Van Nimwegen, "MotEvo: Integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences," *Bioinformatics*, vol. 28, no. 4, pp. 487–494, 2012.
- [93] J. Köhler, S. Philippi, and M. Lange, "SEMEDA: Ontology based semantic integration of biological databases," *Bioinformatics*, vol. 19, no. 18, pp. 2420–2427, 2003.
- [94] J. Pauling, R. Röttger, A. Tauch, V. Azevedo, and J. Baumbach, "CoryneRegNet 6.0 - Updated database content, new analysis methods and novel features focusing on community demands," *Nucleic Acids Res.*, vol. 40, no. D1, pp. 610–614, 2012.
- [95] M. Beckstette, R. Homann, R. Giegerich, and S. Kurtz, "Fast index based algorithms and software for matching position specific scoring matrices.," *BMC Bioinformatics*, vol. 7, p. 389, 2006.
- [96] T. J. Wheeler and S. R. Eddy, "Nhmmer: DNA homology search with profile HMMs," *Bioinformatics*, vol. 29, no. 19, pp. 2487–2489, 2013.
- [97] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.," *Mol. Syst. Biol.*, vol. 7, no. 1, p. 539, 2011.
- [98] J. Koehler, C. Rawlings, P. Verrier, P. Mitchell, A. Skusa, A. Ruegg, and S. Philippi, "Linking experimental results, biological networks and sequence analysis," *Silicio Biol.*, vol. 5, 2004.
- [99] T. Wittkop, D. Emig, S. Lange, S. Rahmann, M. Albrecht, J. H. Morris, S. Böcker, J. Stoye, and J. Baumbach, "Partitioning biological data with transitivity clustering," *Nat. Methods*, vol. 7, no. 6, pp. 419–420, Jun. 2010.
- [100] T. J. Wheeler, J. Clements, and R. D. Finn, "Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models.," *BMC Bioinformatics*, vol. 15, p. 7, 2014.
- [101] J. Baumbach, K. Brinkrolf, L. F. Czaja, S. Rahmann, and A. Tauch, "CoryneRegNet: an ontology-based data warehouse of corynebacterial transcription factors and regulatory networks.," *BMC Genomics*, vol. 7, p. 24, 2006.
- [102] J. Baumbach, K. Brinkrolf, T. Wittkop, and A. Tauch, "CoryneRegNet 2: An Integrative Bioinformatics Approach for Reconstruction and Comparison of Transcriptional Regulatory Networks in Prokaryotes," *J. Integr. Bioinform.*, vol. 3, no. 2, 2006.
- [103] J. Baumbach, T. Wittkop, K. Rademacher, S. Rahmann, K. Brinkrolf, and A. Tauch, "CoryneRegNet 3.0—An interactive systems biology platform for the analysis of gene regulatory networks in corynebacteria and *Escherichia coli*," *J. Biotechnol.*, vol. 129, no. 2, pp. 279–289, Apr. 2007.
- [104] J. Baumbach, "CoryneRegNet 4.0 - A reference database for corynebacterial gene regulatory networks," *BMC Bioinformatics*, vol. 8, p. 429, 2007.
- [105] S. Schaaf and M. Bott, "Target genes and DNA-binding sites of the response regulator PhoR from *Corynebacterium glutamicum*," *J. Bacteriol.*, vol. 189, no. 14, pp. 5002–5011, 2007.
- [106] I. Brune, H. Werner, A. T. Hüser, J. Kalinowski, A. Pühler, and A. Tauch, "The DtxR protein acting as dual transcriptional regulator directs a global regulatory network involved in iron metabolism of *Corynebacterium glutamicum*," *BMC Genomics*, vol. 7, p. 21, 2006.