Integrating Biological Databases in the Context of Transcriptional Regulatory Networks

Rafael Pereira and Rui Mendes

***Abstract*—Several studies show that biological knowledge is growing at a continuous rate and distributed among different databases, making the process of data integration a hard task to perform, because they have different structures, different ways of storing data and also different approaches to export information, and are usually developed to provide information for a specific organism. Due to the large amount of biological data, the process of data integration has been one of the major challenges in the field of bioinformatics as well as discovering information about Transcriptional Regulatory Networks (TRN). When using a single source, this task is not easy to perform since the source often lacks enough information for the successful completion of the task. Therefore it is necessary to find information in several databases in order to create a useful body of knowledge. This work presents a new approach of integrating data related with TRNs for the Escherichia coli by creating a new integrated data repository gathering information from KEGG, EcoCyc, Regulon and NCBI databases.**

***Index Terms*—Data integration, databases, transcriptional regulatory networks and gene.**

I. Introduction

Nowadays one of the main issues addressed in the bioinformatics field is understand the structure and behaviour of complex molecular interaction networks that control cell behaviour.

The vast amount and complexity of biological data retrieved in recent years requires an integrated approach, incentivizing scientists to look for novel approaches in order to address this issue. This task is difficult because researchers often need to retrieve information from several databases that often have different structures.

Currently, the reconstruction of transcriptional regulatory networks requires this sort of integration and is also one of the most important topics in the field of bioinformatics. TRNs are a model of the cellular process that involves many different molecules and is responsible for controlling gene expression. Their study helps us to understand the process of regulatory interactions that link the transcription factors to their target genes. TRNs help us to understand the global organization of regulatory networks and define the functional properties of a cell-type. Furthermore transcriptional regulation is one of the most fundamental mechanisms for controlling the amount of protein produced by cells under different environmental conditions and developmental stages [1].

Integrating data from different sources is a very difficult task within the bioinformatics domain. This task is complex partly because of the term ambiguity in the available databases, terminologies and ontologies.

In the case of TRNs, it is important to integrate several sources that collect information about genes and transcription factors (TF) that are involved in the process of protein regulation.

Thus several databases will be studied in order to retrieve all necessary information in order to perform this reconstruction. However, one often finds that differences between the structure and data types in the bioinformatics databases are a reality.

One way to solve the data integration problem is to create a unique repository that will gather information form these databases.

This work addresses some characteristics about biological data integration, mainly concerning the main biological databases used for regulatory network reconstruction, presents a review of the concepts related to TRNs and the methodology used to develop the integrated data repository.

II. Transcriptional Regulatory Networks

Transcriptional Regulatory Network (TRN) is model of a biological process that occurs within the cells and provides links between genes and their products. In the 1960’s, genetic and biochemical experiments demonstrated the presence of regulatory sequences in the proximity of genes and the existence of proteins that are able to bind those elements and to control the activity of genes by either activation or inhibition of transcription [2].

Research in the Systems Biology field is steadily increasing in the last years and one of the most addressed topics in this area is the simulation of biological systems, whose aim is to perform a reconstruction, in silico and in vivo, of all processes that occur within the cell, both metabolic or regulatory.

The reconstruction of these networks becomes evident and aids the simulations to be paired with experiments, and in fact, they are used by computational scientists to understand the quantitative behaviour of many complex biological systems [3].

In recent years, the exploration of life sciences has been bolstered through the advent of whole genome sequencing. This new information potentiates the reconstruction of genome-scale metabolic networks [4]. But only the reconstruction of metabolic networks is not sufficient to understand the principles about how organisms function.

After this step it is necessary to discover how genetic machinery operates inside an organism, which includes the

Manscript received March 4, 2014; revised May 13, 2014. This work was supported by CNPq, National Counsel of Technological and Scientific Development – Brasil.

The authors are with BioSystems, University of Minho, Braga, Portugal (e-mail: rafatp@gmail.com).

DOI: 10.7763/IJBBB.2014.V4.368

transcriptional regulatory networks.

According to *de-Leon* [5], the regulation is composed by two complementary components; one component is the regulatory genes, e.g. transcription factors and signalling molecules. Transcription factors bind to specific sequences in the DNA and activate or inhibit the transcription of a gene. Signalling molecules carry out the communication between cells and initiate the activation of certain transcription factors in the cells that receive the signal.

The complementary part of these components is the regulatory genome. Every gene contains regulatory sequences that control when and where it is expressed. The regulatory sequences are arranged in units that are termed cis-regulatory modules that contain groups of different transcription factors binding sites as shown in Fig. 1a.

Cis-regulatory behaves as an information processor that reads the regulatory state of the cell and activates or inhibits the gene that it controls. Fig. 1b shows cis–regulatory modules, responsible for activating a gene at a given time and in a specific organism. Activations or inhibitions are rules that form a regulatory network that can be seen in Fig. 1c.

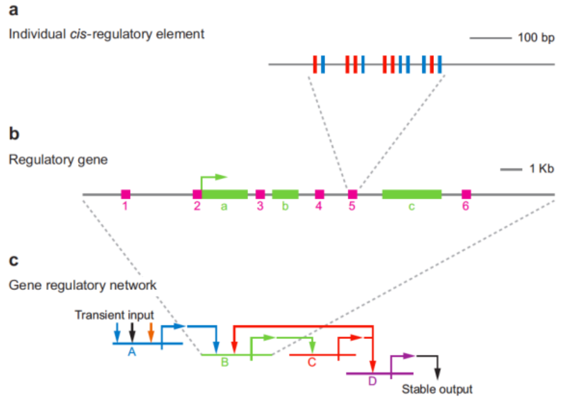

Fig. 1. The gene regulatory hierarchy [5]. (a) Individuals cis-regulatory modules that contain a group of transcription factor binding sites represented through the red and blue boxes. (b) The pink boxes represent the cis-regulatory modules, responsible for expression control and in light green are represented the exons. (c) Represents the inter-regulating, transcription factors and signalling from a small network.

TRNs are usually represented as directed graphs, where each node represents both regulators (i.e., transcription factors) and targets while the edges represent regulatory interactions. This representation can be divided into three levels. The first one includes the set of transcription factors; downstream target genes and the binding sites in the DNA (Fig. 2a.). At the second level, these basic units are arranged into common patterns of interconnections called network motifs (Fig. 2b.). At the third level, motifs are grouped into semi-independent transcriptional units named modules (Fig. 2c.) [6]. Thus at the last level, the regulatory network is composed of interconnecting interactions between the modules that comprise the entire network (Fig. 2d.) [7].

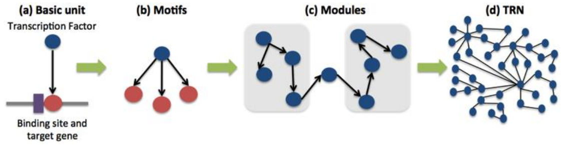

Fig. 2. Comprises the transcription factor, its target gene with DNA recognition site and the regulatory interaction between them. (b) The basic units are organized into networks motifs. (c) Network motifs are interconnected to form semi-independent modules. (d) The whole set of interactions that represent a transcription regulatory network.

In this context, below are described the main concepts related with Transcription Factors (TFs), Motifs, Modules, and finally a point of view about the global organization of TRN.

- Transcription factors are proteins with special abilities and attributes not found in other classes of proteins. Normally they work in pairs or networks to modulate specific regulatory pathways [8], forming multiple interactions that allow for varying levels of control over rates of transcription. TFs have an important function within transcription regulation; they are responsible for recognizing their targets through specific sequences, called motifs [9].
- Motifs represent specific inter-regulation patterns between TFs and target genes [6]. Therefore, a TRN can be defined as a network of motifs, where each these motif has a specific function in determining gene expression [10].
- Modules describe an intermediate level of interaction. Intuitively, one might expect distinct cellular processes to be conveniently regulated by discrete and separable modules [6]. These modules represent a set of motifs that can be interconnected in semi-independent ways. Thus, regulatory networks may be tightly interconnected and several modules can be separated from the rest of the network.
- The Global organization of TRNs is described by parameters derived from graph theory [6]. A TRN may also be represented by a G x R matrix, where G is the number of genes considered and R is the number of transcription factors involved in regulation [9]. The value position (*i*, *j*) is 1 if there is a regulatory relationship between the i -th gene and j -th TF and 0 otherwise. Thus, the most basic elements in these networks are the transcription factors and target genes, and the regulatory interactions between them.

These concepts presented above are fundamental to understand the behaviour of a TRN and also are useful to develop new biotechnical applications as well as to help discovery the nature of diseases. A large amount of information related with TRN can be retrieved from the internet but still not easy to get because in most cases this data is spread over several biological databases such as: NCBI, EcoCyc, KEGG and RegulonDB that will be described in the next section.

III. Biological Databases

Biological databases were developed initially to provide information about genomic sequencing. With the advance of biological techniques and the increase of experimental studies, several databases were created to provide this information for scientific community.

Currently there is a considerable number of biological databases that represent many different kinds of information but this work will be focused on four databases: NCBI, EcoCyc, KEGG and RegulonDB. These databases were chosen specifically because they store information that can used in studies related with TRNs. These databases are described in some detail below:

- **NCBI**: this database arose from one division of NLM

(U.S. National Library Medicine) to provide computational methods that help on biomedical researches. It is the major objective was to develop new information technologies to help understand molecular comprehension and also the controlling of diseases. More specifically the NCBI was created to provide systems that help researchers to store and analyse knowledge from the fields of molecular biology biochemical and genetics, facilitating the use of databases and software in order to retrieve high-level information. NCBI can be considered one of the largest research centers in the field of Bioinformatics. It integrates: literature databases, genomic databases, tools for data mining, tools for sequence analysis and other resources from several biological areas [11].

- **EcoCyc**: this database was created to provide biological information about a specific organism (Eschrerichia coli). The information in this database is mainly about enzymes and metabolic pathways [12]. The data organization of EcoCyc uses a frame knowledge representation system (FRS), that provides an objected-oriented data model, and has several advantages over a database approach because it organizes information within classes—a set of objects that share similar properties and attributes [12]. The main objective of EcoCyc is to store information about proteins, pathways and molecular interaction in *E.coli*. But in recent years its focus was expanded to include annotation and literature-based curation of gene and protein functions of enzymatic, transport and binding reactions, as well as transcriptional regulation, covering the entire genome [13].
- **KEGG**: is a bioinformatics resource for understanding the functions and utilities of cells and organisms from both high-level and genomic perspectives [14]. It provides an integrated resource containing genomic, chemical, and network information while still allowing links to foreign databases. KEGG consists of fifteen main databases, which describe several characteristics like pathway maps, human diseases, organisms and biochemical reactions. The KEGG database can be divided into three levels: the first one called PATHWAY that represents the higher order functions in terms of a network of interacting molecules; the second one called GENES, thatisresponsible of cataloguing genes of bothpartial andcomplete sequences; and the thirdone called LIGAND that stores chemicalcompounds, enzyme molecules and enzymatic reactions [15]. KEGG can be considered a computer representation of a biological system, consisting of molecular building block of genes and proteins (genomic information) and chemical substances (chemical information) that are integrated within a central repository.
- **RegulonDB**: is currently one of the largest databases offering curated knowledge of transcriptional regulatory network for any kind of organism [16]. Initially it was developed to be the main reference of database that offerscurated knowledge of transcriptional regulatory network of *Escherichia coli k-12*. Nowadays, RegulonDB is a relational database service to the scientific community involved in the study of bacteria, offering in an organized and computable form, knowledge on transcriptional regulation that has been manually curated, from original scientific publications. This includes curated information of transcription initiation through the activation or repression of transcription factors (TFs), which bind to individual sites around promoters [17].

TABLE I: DIFFERENCES BETWEEN BIOLOGICAL DATABASES

Database	Knowledge domain	Web Services
NCBI	Literature, genomic, sequence analysis	Yes
EcoCyc	Genes, proteins, pathways and molecular interactions	Yes
KEGG	Pathways maps, human diseases, organisms and biochemical reactions	Yes
RegulonDB	Knowledge on transcriptional regulation	Yes

Table I shows the differences between these biological databases, where can be seen mainly the difference between the knowledge domains.

The choices of databases for this followed the rationale of structure, information that can be retrieved and web-services available for use in order to optimize the gathering process.

IV. Methodology

This proposal involves gathering information, parsing and integration.

Currently, several databases allow users to retrieve information by using Web Services. This section describes a standardized way of integrating systems based on web. Most of these applications use some protocols for communication between the client and the Web Service, like SOAP (Simple Object Access Protocol; www.w3.org/TR/soap). One advantage of Web Services is that client-side applications do not need any intimate knowledge of the database provided by the service itself [18]. Nowadays, scientists are frequently relying on these services for data analysis and several studies have been published using results from these services [19].

This proposal describes a methodology that uses the Web Services available for each database described in the previous section, in order to retrieve information about genes from a specific organism, the

Escherichia coli K-12. In order to achieve this, it was necessary to find a list of gene identifiers for each database, because there is still no standard that can be used for these databases.

It was necessary to create a file for each database containing a list of gene identifiers from *E.coli* since identifiers are database specific. A Java class was implemented to read these files and connect to the Web Services, and then it returns all information available for these genes. It is important to remember that each database has a different structure to store the information, and also different output formats. For instance, RegulonDB and KEGG return a text file, NCBI and EcoCyc a XML (Extended Markup Language; www.w3.org/XML) file, thus increasing the complexity of the process of integration.

After the data retrieval process, it was necessary to perform an analysis over the structure of each file and develop a parser to transform the information into XML files since this enables users to perform searches and also to easily navigate across the information thus facilitating the integration process. Storing information into XML files is currently a valid alternative over using relational databases. It is becoming a standard way of representing data changes, a form of web publication and also a flexible syntax that allows the same information to be represented in different ways. Fig. 3 depicts the process of data extraction and publication into XML files.

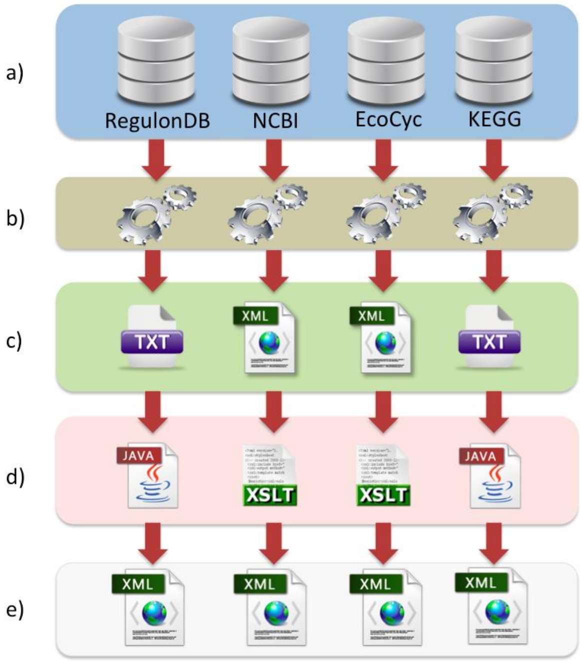

Fig. 3. The basic structure of retrieving and parsing data into a standard format. (a) The databases that compose the system; (b) the web services used to connect and retrieve the data; (c) the output file formats from the web services; (d) the parsers that were developed to put this data into XML format.

Once all data is in the same format, the structure of each file was analyzed in order to create a unique repository able to store all information from these databases. Before data integration can start, it is necessary to perform data validation. In order to perform this, XQuery[1]— a language used to perform queries in XML documents — was used. During the validation process, some issues began to appear such as the number of annotated genes in each database, the difference between gene identifiers, a variety of names recorded for the same gene, some genes are found in a database and not in others, and so on. Since there is an apparent incompatibility between these databases, it was important to search for common features among these files. For this specific organism, there is an attribute that is common to all databases, called *B-number* created by Blattner, that is commonly used as a gene identifier [20].

Fig. 4 shows the relation between number of genes and *B-numbers* associated in each database. As can be seen, not all genes have a *B-number*. In two databases, almost all genes have a B-number; in NCBI that has 4498 genes, 4496 have *B-numbers* and in KEGG all genes are associated with a *B-number*.

In the other two databases this inconsistency is more visible. For instance, EcoCyc has a total of 4501 genes, where 230 of these does not have *B-numbers* associated. In RegulonDB only 4496 out of 4639 genes have a *B-number*. Taking this into account, these databases present several incompatibilities related with this information, making the integration process a very difficult task to perform.

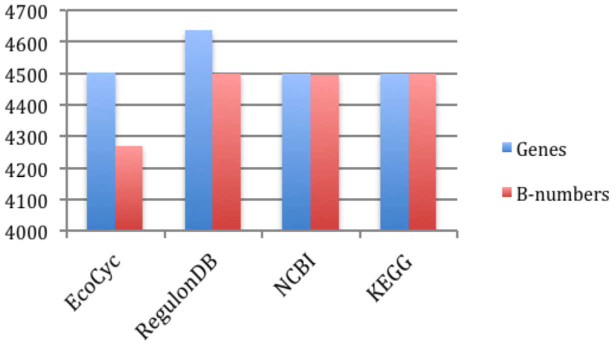

Fig. 4. Comparison between number of genes and B-numbers associated in each database.

When contemplating data integration, there are other problems besides this one, as can be seen in Fig. 5 that shows the difference between the number of genes found in these databases. The large number of coincidences occurs when comparing KEGG and NCBI since they both have the same number of genes. is the same happens when comparing the EcoCyc database and RegulonDB, since only three genes that are present in EcoCyc do not appear in RegulonDB.

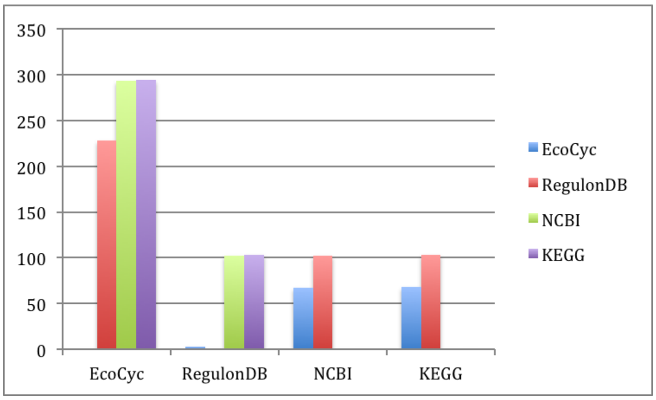

Fig. 5. Difference between the number of genes found in each database.

The worst case is when comparing KEGG and EcoCyc since 294 genes are present in KEGG don’t seem to be present in EcoCyc. It is necessary to gather a large amount of information to perform data integration in TRNs. In order to tackle the problems that arise due to the inconsistency between these databases, an integration process was developed. This involves gathering all information related with TRNs available among these databases and storing this data in a unique repository, that was called the KREN as can be seen in Fig. 6.

The KREN repository was developed to store information retrieved from the previously databases described. Each XML file was analyzed and a search by information patterns between these files was performed. Finally, some common characteristics were found allowing us to build the repository. Furthermore a reader for each XML file was developed in order to load the information into the KREN structure.

The main component of this repository is the gene. All information inside this data source is related with this entity and all genes are identified by a unique *B-number* that is common among the databases. This information was validated across all databases.

[1] http://www.w3.org/TR/xquery/

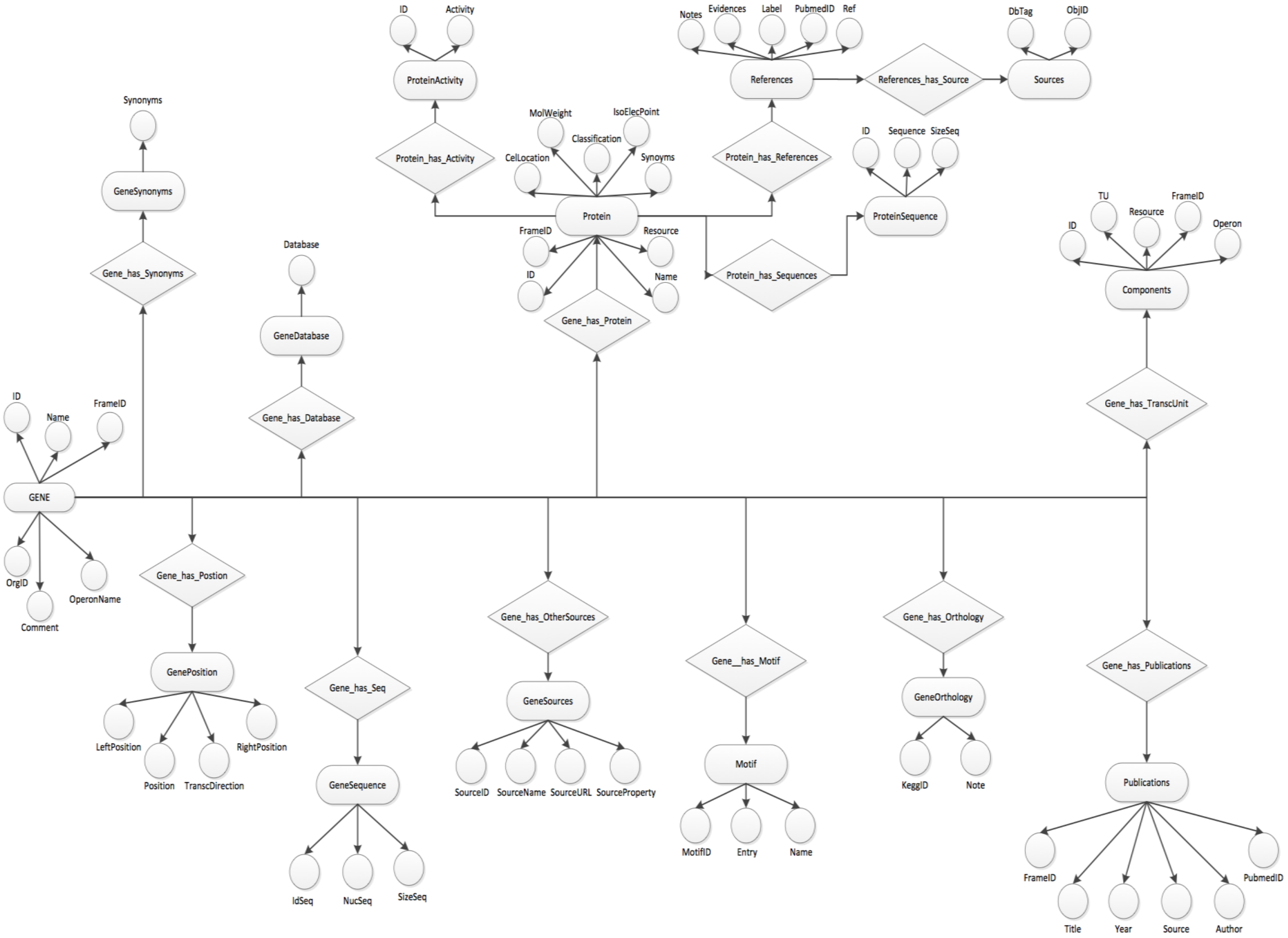

Fig. 6. The KREN repository.

V. Conclusion

The traditional approach to retrieve information from biological databases has been one of hottest topics in this area. Currently, it is getting more difficult since there are a large number of new heterogeneous databases, often sporting different structures, formats, and ways of storing information as well as different ways of providing this information to users. These factors along with the dispersion of information entail the fact that biological data integration process is a very hard task to perform.

This work addresses the process of biological data integration in the field of TRNs, bringing an alternative approach to retrieve, integrate and store information from some of more relevant databases. Nowadays, this repository contains data about TRNs from *Escherichia coli strain K-12 MG1655*. It can easily be extended in order to incorporate data from other organisms that are available in the databases cited in this paper.

References

[1] I. Simon, J. Barnett, N. Hannett *et al.*, "Serial regulation of transcriptional regulators in the yeast cell cycle," *Cell*, vol. 106, no. 6, pp. 697–708, Sep. 2001.

[2] T. Schlitt and A. Brazma, "Current approaches to gene regulatory network modelling," *BMC bioinformatics*, vol. 8, Suppl 6, p. S9, Jan. 2007.

[3] J. Sanz, J. Navarro, A. Arbués *et al.*, "The transcriptional regulatory network of Mycobacterium tuberculosis," *PloS one*, vol. 6, no. 7, p. e22178, 2011.

[4] C. L. Barrett and B. O. Palsson, "Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach," *PloS Computational Biology*, vol. 2, no. 5, p. e52, May 2006.

[5] S. B. T. de Leon and E. H. Davidson, "Gene regulation: gene control network in development," *Annual Review of Biophysics and Biomolecular Structure*, vol. 36. p. 191, Jan. 2007.

[6] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann, "Structure and evolution of transcriptional regulatory networks," *Current Opinion in Structural Biology*, vol. 14, no. 3, p. 283, 291, 2004.

[7] J. Carrera, G. Rodrigo, A. Jaramillo, and S. F. Elena, "Reverse-engineering the Arabidopsis thaliana transcriptional network under changing environmental conditions," *Genome Biology*, vol. 10, no. 9, p. R96, Jan. 2009.

[8] D. Yusuf, S. L. Butland, M. I. Swanson *et al.*, "The transcription factor encyclopedia," *Genome Biology*, vol. 13, no. 3, p. R24, Jan. 2012.

[9] N. Sun and H. Zhao, "Reconstructing transcriptional regulatory networks through genomics data," *Statistical Methods in Medical Research*, vol. 18, no. 6, pp. 595–617, Dec. 2009.

[10] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in t the transcriptional regulation network of Escherichia coli," *Nature Genetics*, vol. 31, no. 1, pp. 64–68, May 2002.

[11] National Libreary of Medicine. (Jan. 2014) National Center for Biotechnology Information. [Online]. Available: http://www.ncbi.nlm.nih.gov.

[12] P. Karp and M. Riley, "Ecocyc: The resource and the lessons learned," *Bioinformatics: Databases and Systems*, pp. 1–13, 2002.

[13] I. M. Keseler, J. Collado-Vides, S. Gama-Castro *et al.*, "EcoCyc: a comprehensive database resource for Escherichia coli," *Nucleic Acids Research*, vol. 33, pp. D334–D337, Jan. 2005.

[14] M. Tanabe and M. Kanehisa, *Using the KEGG Database Resource*, Chapter 1, June 2012.

[15] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[16] H. Salgado, S. Gama-Castro, M. Peralta-Gil *et al.*, "RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network,

operon organization, and growth conditions," *Nucleic Acids Research*, vol. 34, pp. D394–D397, Jan. 2006.

[17] H. Salgado, A. Santos, U. Garza-Ramos *et al*., "RegulonDB (version 2.0): a database on transcriptional regulation in Escherichia coli," *Nucleic Acids Research*, vol. 27, no. 1, pp. 59–60, Jan. 1999.

[18] T. Cokelaer, D. Pultz, L. M. Harder, J. Serra-Musach, and J. Saez-Rodriguez, "BioServices: a common Python package to access biological Web Services programmatically," *Bioinformatics,* vol. 29, no. 24, pp. 3241–3242, Dec. 2013.

[19] S. Schultheiss, M. Munch, G. Andreeva, and G. Rätsch, "Persistence and availability of Web services in computational biology," *PloS One*, vol. 6, no. 9, p. e24914. Jan. 2011.

[20] F. R. Blattner, "The complete genome sequence of escherichia coli K-12," *Science*, vol. 277, no. 5331. pp. 1453–1462, Sep. 1997.

Rafael Pereira has received his bachelor degree in computer science at Centro Universitário Franciscano, Santa Maria/ Brasil; M.Sc. in bioinformatics at University of Minho, Braga/Portugal; Currently he is a Ph.D student in informatics and member of researches groups in: bioinformatics and systems biology interdisciplinary initiative (University of Minho); computational physics, complex systems, genetic theory (Federal University of Santa Maria); programming languages and data bases (Federal University of Santa Maria). Its main areas of activity are reconstruction of genome-scale regulatory models, databases, strain optimization and metabolic engineering. It is part of ReGeReM's porject, which is a natural follow up on the previous work done by research team within the field of Bioinformatics and Systems Biology. It addresses a very relevant research problem in this field that has important applications in Biotechnology: the development of algorithms and computational tools for reconstruction of genome-scale regulatory models for microorganisms.

Rui Mendes has received his Ph.D in artificial intelligence at Univeristy of Minho and an assistant professor in the same university. His research interests include evolutionary computation and multi-agent systems.