Reconstructing Transcriptional Regulatory Networks Using Data Integration and Text Mining

Rafael T. Pereira*, Hugo Costa†, Sónia Carneiro*†, Miguel Rocha* and Rui Mendes*
*Centre of Biological Engineering
University of Minho, Braga, Portugal 4710-057
Email: rafatp@di.uminho.pt, soniacarneiro@deb.uminho.pt, mrocha@di.uminho.pt, rcm@di.uminho.pt
†SilicoLife, Braga, Portugal
Email: hcosta@silicolife.com

***Abstract*—Transcriptional Regulatory Networks (TRNs) are powerful tool for representing several interactions that occur within a cell. Recent studies have provided information to help researchers in the tasks of building and understanding these networks. One of the major sources of information to build TRNs is biomedical literature. However, due to the rapidly increasing number of scientific papers, it is quite difficult to analyse the large amount of papers that have been published about this subject. This fact has heightened the importance of Biomedical Text Mining approaches in this task. Also, owing to the lack of adequate standards, as the number of databases increases, several inconsistencies concerning gene and protein names and identifiers are common. In this work, we developed an integrated approach for the reconstruction of TRNs that retrieve the relevant information from important biological databases and insert it into a unique repository, named KREN. Also, we applied text mining techniques over this integrated repository to build TRNs.**

However, was necessary to create a dictionary of names and synonyms associated with these entities and also develop an approach that retrieves all the abstracts from the related scientific papers stored on PubMed, in order to create a corpora of data about genes. Furthermore, these tasks were integrated into @Note, a software system that allows to use some methods from the Biomedical Text Mining field, including an algorithms for Named Entity Recognition (NER), extraction of all relevant terms from publication abstracts, extraction relationships between biological entities (genes, proteins and transcription factors). And finally, extended this tool to allow the reconstruction Transcriptional Regulatory Networks through using scientific literature.

I. Introduction

Research in the Systems Biology field is steadily increasing in the last years and one of the most addressed topics is the modeling and simulation of biological systems, whose aim is to recapitulate, *in silico* and *in vivo*, all processes that occur within the cell, both metabolic and regulatory. All information that is necessary to perform this kind of simulation should, in principle, be available in one of the numerous biological databases. Indeed, in recent years, a large number of biological databases have appeared, which can be divided according to their characteristics, the type of information that is stored and also if they are specific for some organism. This entails that there is an increasing amount of biological information that can be retrieved from these databases. Also, there is an increase in the complexity of integrating this knowledge, motivating researchers to look for novel methods to address this problem.

Besides the increasing number of databases in the field of biological sciences, there is also a vast amount of scientific publications, that is steadily increasing, both on journals and books. This is illustrated by the growth of the PubMed[1] database, where there are currently more than 24 million citations. Through this database, it is possible to find scientific papers from many life science areas concerning diverse subjects. However, finding information concerning a given subject in PubMed can be a daunting task due to the vast number of papers related to any specific search.

This paper will present an integrated approach for the reconstruction of regulatory networks by using text mining. Furthermore, an integrated repository was developed, allowing users to search for all necessary information in a unique source, thus saving time and avoiding the frequent inconsistencies that arise when one needs to work with different sources.

As a case study, these methods will be applied to reconstruct several small networks from the bacterium *Escherichia coli*, substrain *K-12 MG1665*, and validate it through scientific experiments referred in the literature.

II. Transcriptional Regulatory Networks

Transcriptional Regulatory Networks (TRN) are models of a biological process that occurs within the cells and provide links between genes and their products. In the 1960s, 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 transcription activation or inhibition [1].

The structure of TRNs may be described by a series of components:

Transcription Factors (TFs) are proteins that either promote or block the transcription in specific regulatory pathways, they may function alone or with other proteins in a complex [2];

Motifs represent specific patterns of inter-regulation between TFs and their target genes, their primary function is to determine the gene expression [3];

Modules are an intermediate level of interaction, representing a set of Motifs that can be interconnected in semi-independent ways [4]

[1]PubMed database: http://www.ncbi.nlm.nih.gov/pubmed

978-1-4673-6799-8/15/$31.00 ©2015 IEEE

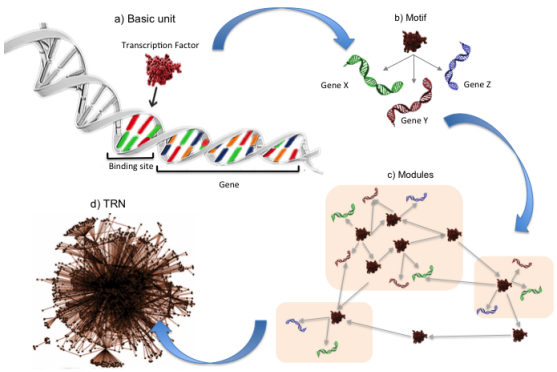

Fig. 1: TRN organization from basic units until the transcriptional regulatory network. a) Basic unit, contains the Transcription Factor, its target gene with DNA recognition site and the regulatory interaction between them. b) The basic units are organized into networks motifs. c) Network motifs are interconnected to form semi-independent modules. d) The whole set of interactions that represent a transcription regulatory network.

III. Biological Databases

Biological databases were developed initially to provide information about genomic sequencing. With the advance of biological techniques and the increase in the number of experimental studies, more databases were created to provide this information for the scientific community. There is currently a large number of biological databases that provide information about several aspects of biological information. This work will focus on four databases/ repositories that offer information that is necessary for this work: NCBI, EcoCyc, KEGG and RegulonDB. Those were chosen specifically because they store information that can be used in studies related with TRNs. A summary of each one is provided below:

NCBI This repository includes a set of databases and tools which arose from one division of NLM (U.S. National Library Medicine) to provide computational methods for biomedical researchers. Its major objective was to develop new information technologies to help understand molecular comprehension and also the controlling of diseases. More specifically, NCBI was created to provide systems that help researchers to store and analyse knowledge from the fields of molecular biology biochemistry and genetics, facilitating the use of databases and software to retrieve high-level information [5].

RegulonDB This is currently one of the largest databases offering curated knowledge about transcriptional regulatory networks for any kind of organism [6]. Initially, it was developed to be the reference database offering curated knowledge about the 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, manually curated knowledge about transcriptional regulation gathered from original scientific publications [7].

EcoCyc This database was created to provide biological information about a specific organism (*Escherichia coli*). The information in this database is mainly about enzymes and metabolic pathways. 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 each representing a set of objects that share similar properties and attributes [8]. The main purpose 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 [9].

KEGG It is a bioinformatics resource for understanding the functions and utilities of cells and organisms from both high-level and genomic perspectives [10]. 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.

Each database offers information using a different approach; there are different means of accessing and extracting information; files retrieved are from different types; there is an inconsistency in both the number of genes and proteins from each of these sources and so on. Thus, in order to address these problems, one needs an integrated solution. The next section describes the KREN repository, that allows to store the information retrieved from these databases in an unique source.

IV. KREN, an integrated repository for gene and proteins information

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.

Currently, several databases allow users to retrieve information by using Web Services [11]. Most of these applications use some protocols for communication between the client and the Web Service, like SOAP (Simple Object Access Protocol)[2]. An advantage of Web Services is that client-side applications do not need any intimate knowledge of the database behind the service itself [12].

It is essential for this work to gather as much information as possible related with TRNs, mainly about genes, proteins, transcription units and bibliography references to perform the reconstruction of these networks. Thus, to solve the data integration problem it was necessary to create a repository that compiles all the information gathered from these databases.

The KREN repository [13] was developed to store information retrieved from the four databases that were previously described. The main component of this repository is the gene, shown in Figure 2.

[2]Simple Object Access Protocol - http://www.w3.org/TR/soap

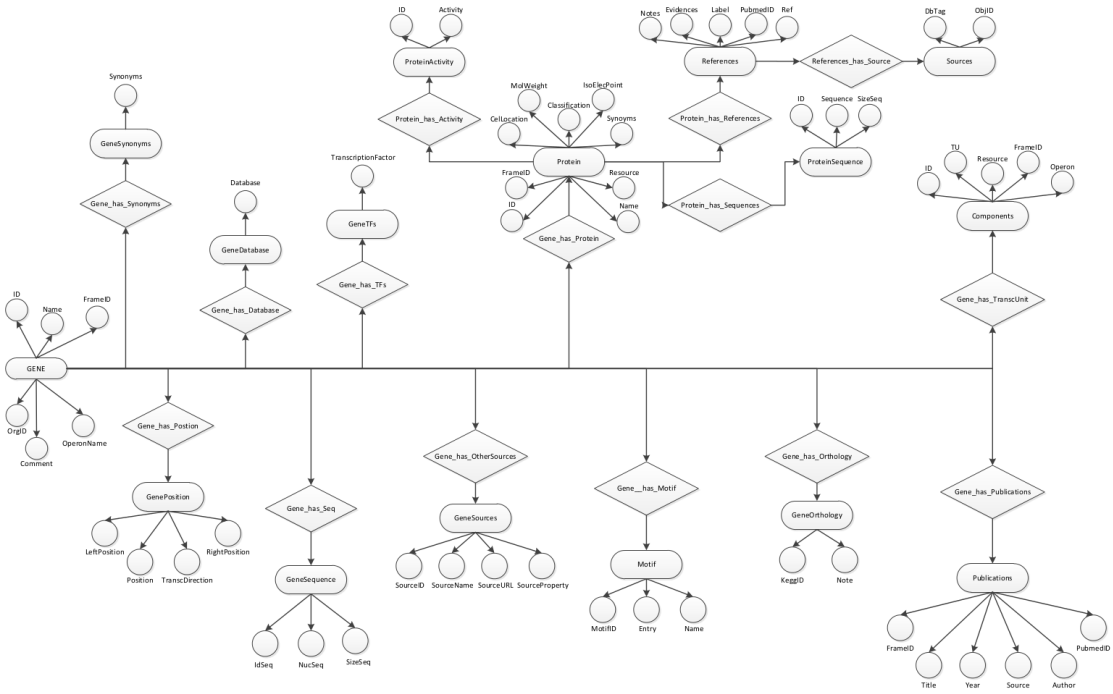

Fig. 2: KREN repository

For this work, the KREN was used to provide some specific information, such as gene names, protein names, gene synonyms, protein synonyms, transcription factors and also a list of identifiers from the PubMed database, where it is possible to retrieve all scientific publications related with genes from *E. coli*. 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.

Using this repository, it is possible to retrieve all PubMed identifiers, name, synonyms and proteins associated with any given gene. This information will be useful to create a Biomedical Text Mining (BioTM) resource like corpus and lexical resources (dictionaries) that will be needed to accomplish the goal of reconstructing regulatory networks.

V. Text mining applied to biological literature

It is currently very hard to keep track of the increase of scientific publications in the biological field, which are currently spread across several literature sources. For instance, if one performs a simple query on PubMed for a gene name as "*rpoA*" from *Escherichia coli*, the results will include around 330 matching publications. Sifting through all these papers is a time consuming task that will probably yield a small dividend. Thus, text mining tools can help simplify this daunting task by performing an automatic search that will highlight a smaller subset of papers whose contents will hopefully be more interesting for a given researcher.

Text mining aims to extract high-quality information from text that is deemed important for a specific case study. These applications tend to be an automated way of searching and extracting information over large sets of natural language texts. Most biomedical text mining systems contain components that allow the recognition of biological entities or concepts in text, as well as relations between these entities [14]. BioTM can be divided in two main areas: Information Retrieval (IR) and Information Extraction (IE). Whereas IR allows the extraction of all the parts from a document (e.g., abstract) that is stored on a repository; IE can be divided in Named Entity Recognition (NER), where biological entities can be recognized in the text and Relation Extraction (RE) that finds relations between these entities.

An important part of BioTM are computational methods from the Natural Language Processing (NLP) field that are used to retrieve useful information from natural language sources and have been applied to recover information, text mining and voice recognition. Essentially, NLP attempts to deal with two parts of textual analysis: the structure and the meaning. The first one can be divided between the morphological (which is the comprehension of words) and the syntactical analysis (defines the structure in the phrase based on the relationship among words). Finally, semantic analysis, associates meaning to the syntactic structure.

Some tools already implement some of these features like GATE[3] and StanfordNLP[4]. Both perform syntactic and semantic annotation and also NER. The last one aims to search and classify terms in a collection of texts, for instance all of occurrence of the gene and proteins names within a set of scientific papers. This task is not easy to perform because there is no comprehensive dictionary of gene and protein names or biological entities [1].

Over the last few years, the Biosystems research group

[3]available on https://gate.ac.uk
[4]available on http://nlp.stanford.edu/

at the University of Minho and the SilicoLife company have worked together in the BioTM field. In this period, a software platform for BioTM called @Note[5], was developed. A major reformulation, including the development of several novel features has been implemented leading to its current upgraded version (2.0). @Note was developed in Java and uses a MySQL database, which copes with the most important IR and IE tasks and promotes multi-disciplinary research. The main goals of this framework are to facilitate the processes of curation and literature annotation; to use already developed models to automate tasks like text annotation and document retrieval; to configure and use models without any need for programming; to translate and validate models and finally to allow developers to extend the application to provide new functionalities [15].

@Note2 will be the core of this work, providing many functionalities like its Application Programming Interface (API) for lexical resource creation, Corpora Management and text annotation as entities and relations, that will be described in more detail on section VI.

A. Discovery of transcriptional regulatory networks by using text mining

Identifying relationships between genes and proteins from scientific papers is still a very difficult task to perform, mainly because it is necessary to recognize and categorize several types of biological events, like: positive/negative regulation, binding, coding, over-expressing, activation/deactivation and so on [16], as well as gene names and proteins. This is illustrated in Figure 3 that shows a gene and a protein name and the verbs that can be used to infer the semantic interaction between this pair of entities.

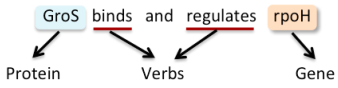

Fig. 3: Text fragment that illustrates a relationship between protein GroS and the rpoH gene. The verbs "binds" and "regulates" represents a regulatory event.

Extracting these events from the literature is still a hard task that initially involves recognizing the named entities that are involved. With the parallel evolution of the molecular biology field in many research groups around the world there is a huge variation on the nomenclature and synonyms from genes and proteins. There is still no agreed standard for these names. For instance, it is common to find gene names that are the same as their product. Besides this task, extracting information about regulatory interactions is complicated because the regulation processes can be expressed in several ways, such as a biological process and not only as a simple interaction between two entities, that could be more obvious. Thus, this process requires an improved approach to deal with these issues.

The next section will propose a workflow where some techniques and methods will be applied to create a new approach using text mining for the identification and extraction of TRNs from the literature.

[5]available on http://www.anote-project.org

VI. Methods

An approach is suggested to deal with complex tasks related with the reconstruction of regulatory networks. This is performed through integrating data from different sources, retrieving scientific publications from PubMed and finally applying Biomedical Text Mining.

First of all, it was necessary to choose an organism to retrieve the information related with TRNs. Since the *Escherichia coli* bacterium is the most studied organism, thus also known as a gold standard, it was the obvious candidate. For this specific approach, the *Escherichia coli K-12* substrain *MG1665* was chosen.

To retrieve information about this organism, the KREN repository was used. It stores several types of data like gene names, protein names, synonyms and PubMed identifiers that are deemed important to retrieve all the literature concepts that can be used in this work.

As it was presented in Section IV, this source of information is very useful for this work, because it integrates data from four major databases available on the web: NCBI, RegulonDB, EcoCyc and KEGG. Hence, all information related with genes and proteins was stored in an unique repository. Initially, the data extracted from KREN will be: the gene names, gene synonyms, protein names, protein synonyms and also PubMed identifiers.

A workflow is shown in Figure 4, where it is possible to see all the steps for the proposed approach, ranging from the data integration process until the reconstruction of the TRNs.

A. @Note Software System

This framework is used in this work because it implements several functionalities that will be useful in this work such as algorithms for NER, building dictionaries and the creation of corpora.

1) Information Retrieval - PubMed Search and Corpora Creation: PubMed is the most important database for scientific literature and also provides a search engine that allows retrieving information through the use of web services.

In this step, a search was implemented to extract all scientific papers from PubMed that are related strictly with *E. coli K-12*. This search is based on KREN, where it is possible to get a list of PubMed identifiers for each gene from the distinct sources. As a result of this task, all these papers will be stored on the @Note database, indexed by gene identifier, thereafter creating a corpora.

2) Creation of a dictionary: To build a dictionary, it is essential to run a NER dictionary based algorithm. A dictionary is created using the KREN repository to save gene and protein terms and synonyms for a specific organism (in this case study, *Escherichia coli*). The KREN already has names and synonyms for proteins and genes.

Once this information is stored on the @Note database, the next step is to build the corpora.

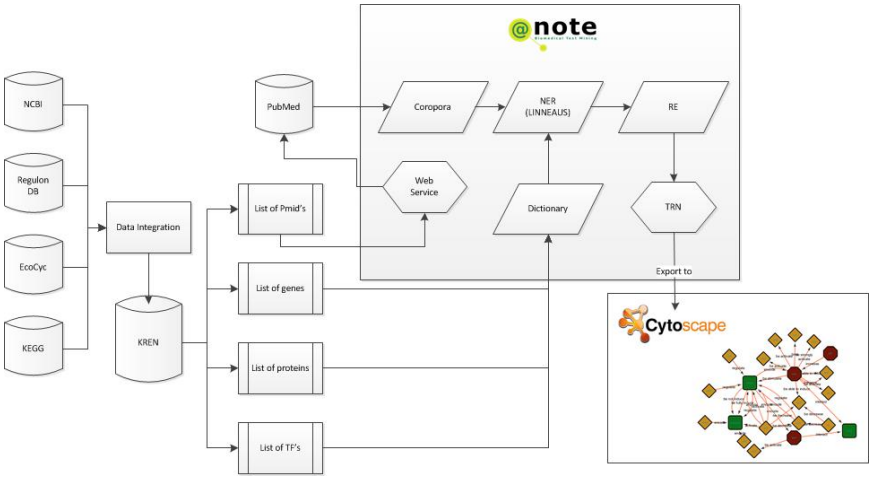

Fig. 4: Workflow representation for the task sequence to reconstruct a TRN.

3) Named Entity Recognition: The task of identifying names, synonyms, expressions may be performed using NER. Nowadays, it is possible to find several algorithms for performing this task. For this work, we chose to use the LINNAEUS algorithm [17][6]. Firstly, this system was developed to recognize names of organisms automatically, but it has evolved to an algorithm for general purpose NER (e.g., genes and proteins). One of the most interesting features of this approach is the use of regular expression based lexicon resources (i.e., dictionaries) for entity identification, thus providing a convenient way of finding and identifying entities.

In order to decrease some issues that may appear when the NER process is performed, a series of preprocessing steps is performed, first by removing stop-words or using a part of speech tagging mechanism (POS-tagging), which is a process that labels words according to their grammatical features, e.g.: verbs, nouns and adjectives.

In the next step, it is necessary to combine the corpora and the dictionary created to apply the NER. This method is implemented on @Note and uses the LINNEAUS algorithm to identify the entities.

4) Relation Extraction: Even though it is planned in the pipeline, the ability of extracting relations from text about biological events is still not fully implemented. Automating this process has been a challenge in the BioTM area [18]. There are several attempts focusing mainly in the automatic recognition, in the normalization, and on mapping these biological entities [19], however some advanced approaches must be developed.

The objective in this step is to implement a linguistic based approach to identify these interactions, using advanced methods of NLP such as Shallow [20], Deep processing [21] or Dependency parsing [22]. The starting point will be to make a list of possible triggers (verbs that will be essential to determine an event, e.g.: regulation, binding, coding, etc.). To perform the information extraction, this approach uses the phrase as a basic unit to analyse and verify the existence of any meaning or relation that could be identified.

The next step, involves the extraction and characterization of biological relationships, that is composed by three main steps: build a syntactic layer based on the textual layer; match the syntactic layer with the semantic layer; and, finally, the creation of rules to extract and characterize the relationships using the verb to characterize an interaction.

However, the number of relationships retrieved from a phrase is proportional to the number of identified clues. An example: assuming a determined phrase, containing five biological entities and three verbal groupings, the model of extraction for each verbal grouping defines the entities that are located upstream and downstream, based on the verbal relative position. The biological entities that comprise a relationship are delimited upstream by the previous verbal grouping (VG) or by the beginning of the phrase and downstream by the verbal grouping immediately next to it or by the ending of the phrase.

All these methods are implemented on @Note framework, that allows extracting this type of relationships in an automatically way. Beside this, it was also performed an export process that allows network visualization on Cytoscape [7].

The preliminary results which will be shown in this section were performed based on the regulatory model *iMC1010* published in 2004 by Covert [23]. In order to build a new regulatory model from *E. coli*, it is necessary to analyse several Boolean rules that determine when a gene will be activated or not. The model published by Covert et al provide a set of these rules, each one of them associated to a unique gene. As an example, the rule to activate the gene *sodA* is given bellow:

sodA: (NOT(ArcA OR Fur) OR (MarA OR Rob OR SoxS))

Thus, for reconstructing this TRN, it is necessary to create a network for each gene and then apply a merge process between these networks. The result is shown in Figure 5.

Thus, it is possible to conclude if the network reconstructed follows the rule defined in the previous model.

Although these networks apparently present certain consistency, it is possible to find an exception, illustrated in Figure

[6]Linneaus Project - http://linnaeus.sourceforge.net/

[7]Cytoscape - http:\\cytoscape.org

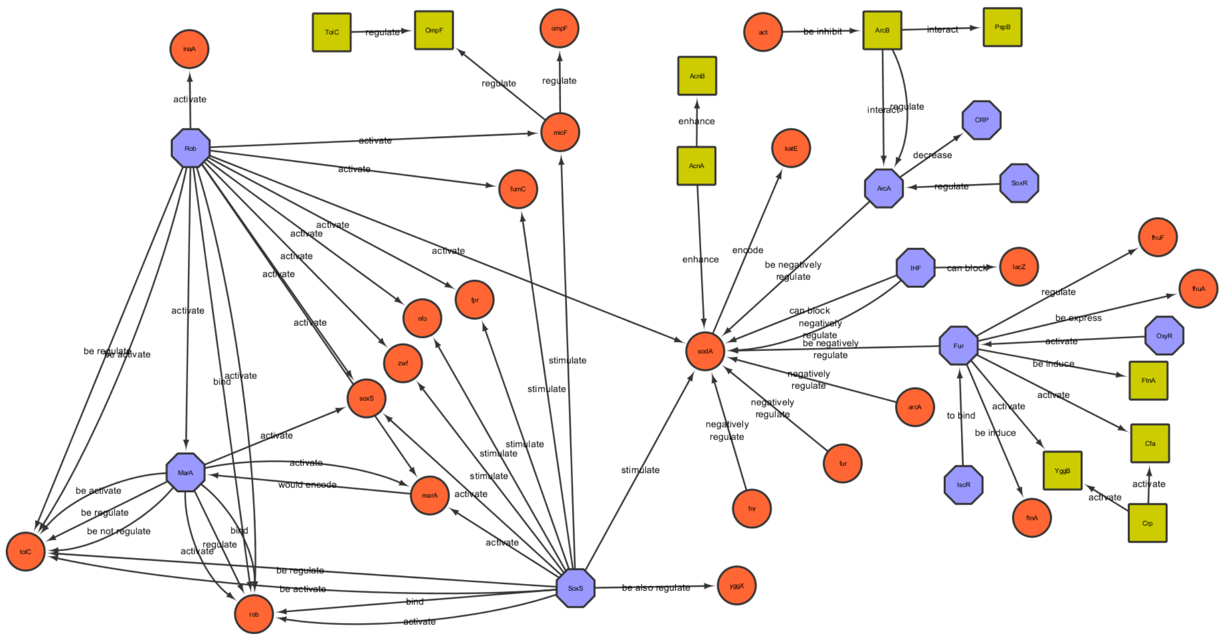

Fig. 5: Small part of Transcrptional Regulatory Network from *Escherichia coli K-12* MG1665 reconstructed by this approach

??, where the edge between the nodes *tolC* and *MarA* is diverging. It occurs because one type of relation was identified: two of them represent an activating action and the other an inhibiting action. It is resulted of the sentence extracted from an abstract of a paper published by Zhang et. al [24] which states: ”Two previously identified *tolC* promoters, *p1* and *p2*, are not regulated by *MarA*,...”. In this sentence, the specific promoters do not allow the gene *tolC* to be regulated by these transcription factors; on the other hand, this gene could be regulated by another promoters, as may be the case.

VII. Conclusion

Recently, the rapid increase of scientific papers augmented the importance of using tools for data integration and fostered the use of text mining methods. It is possible to find several types of biological information dispersed among different databases like experimental data, gene sequencing, bibliography or chemical compounds. Given the fact that this information typically does not conform to any standards, a data integration approach is paramount to gather all information necessary for any given task.

For this work, the KREN was used to provide some specific information and also a list of identifiers from the PubMed database, where it is possible to retrieve all scientific publications related with genes from *E.coli*. Indeed, an approach was developed to build TRNs adding new functionalities for an existent framework called by @Note.

The results presented in this work are quite promising, however a deep process of validation in order to compare with the regulatory model existent is still necessary.

Acknowledgment

This work was supported by grant from the CNPq - ”National Council of Technological and Scientific Development, Brasil”.

References

[1] R. T.-H. Tsai *et al.*, “NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition.” *BMC bioinformatics*, vol. 7 Suppl 5, p. S11, 2006.

[2] D. Yusuf *et al.*, “The Transcription Factor Encyclopedia,” *Genome Biology*, vol. 13, no. 3, p. R24, 2012. [Online]. Available: http://dx.doi.org/10.1186/gb-2012-13-3-r24

[3] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, “Network motifs in the transcriptional regulation network of Escherichia coli.” *Nature genetics*, vol. 31, no. April, pp. 64–68, 2002.

[4] 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, pp. 283–291, 2004.

[5] U.S. National Library of Medicine, “National Center for Biotechnology Information,” http://www.ncbi.nlm.nih.gov/, 1988, online; accessed 10-December-2012].

[6] H. Salgado *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, 2006.

[7] H. Salgado, A. Santos, U. Garza-ramos, and J. E. Helden, “RegulonDB (version 2 . 0): a database on transcriptional regulation in Escherichia coli,” *Nucleic acids research*, vol. 27, no. 1, pp. 59–60, 1999.

[8] P. D. Karp, M. Riley, S. M. Paley, a. Pellegrini-Toole, and M. Krummenacker, “EcoCyc: Encyclopedia of E. coli Genes and Metabolism,” *Nucleic Acids Research*, vol. 25, no. 1, pp. 43–50, 1997.

[9] I. M. Keseler *et al.*, “EcoCyc: A comprehensive database resource for Escherichia coli,” *Nucleic Acids Research*, vol. 33, pp. 334–337, 2005.

[10] M. Tanabe and M. Kanehisa, “Using the KEGG database resource,” *Current Protocols in Bioinformatics*, pp. 1–54, 2012.

[11] P. B. T. Neerincx and J. A. M. Leunissen, "Evolution of web services in bioinformatics," *Briefings in bioinformatics*, vol. 6, no. 2, pp. 178–188, 2005.

[12] 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 (Oxford, England)*, vol. 29, no. 24, pp. 3241–2, 2013. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/early/2013/09/22/bioinformatics.btt547

[13] R. Pereira and R. Mendes, "Integrating biological databases in the context of transcriptional regulatory networks," *International Journal of Bioscience, Biochemestry and Bioinformatics*, vol. 4, no. 5, pp. 345–350, 2013.

[14] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining." *Brief Bioinform*, vol. 6, no. 1, pp. 57–71, 2005.

[15] A. Lourenço *et al.*, "@Note: A workbench for Biomedical Text Mining," *Journal of Biomedical Informatics*, vol. 42, no. 4, pp. 710–720, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2009.04.002

[16] S. Ananiadou, P. Thompson, R. Nawaz, J. McNaught, and D. B. Kell, "Event-based text mining for biology and functional genomics." *Briefings in functional genomics*, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24907365

[17] M. Gerner, G. Nenadic, and C. M. Bergman, "LINNAEUS: a species name identification system for biomedical literature." *BMC bioinformatics*, vol. 11, p. 85, 2010.

[18] C. Friedman, P. Kra, H. Yu, and A. Rzhetsky, "GENIES : a natural-language processing system journal articles," *Bioinformatics*, vol. 17, 2001.

[19] J. M. Temkin and M. R. Gilder, "Extraction of protein interaction information from unstructured text using a context-free grammar," *Bioinformatics*, vol. 19, no. 16, pp. 2046–2053, Oct. 2003. [Online]. Available: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btg279

[20] G. Neumann and J. Piskorski, "A shallow text processing core engine," *Computational Intelligence*, vol. 18, pp. 451–476, 2002.

[21] B. Crysmann *et al.*, "An integrated architecture for shallow and deep processing," in *University of Pennsylvania*, 2002, pp. 441–448.

[22] S. Kubler, R. McDonald, J. Nivre, and G. Hirst, *Dependency Parsing*. Morgan and Claypool Publishers, 2009.

[23] M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson, "Integrating high-throughput and computational data elucidates bacterial networks," *Nature*, vol. 429, no. 6987, pp. 92–96, May 2004. [Online]. Available: http://dx.doi.org/10.1038/nature02456http://www.nature.com/nature/journal/v429/n6987/suppinfo/nature02456_S1.html

[24] A. Zhang, J. L. Rosner, and R. G. Martin, "Transcriptional activation by MarA, SoxS and Rob of two tolC promoters using one binding site: a complex promoter configuration for tolC in Escherichia coli." *Molecular microbiology*, vol. 69, no. 6, pp. 1450–5, Sep. 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2574956\&tool=pmcentrez\&rendertype=abstract