GREAT: Gene Regulation EvAluation Tool

Catia Machado[1], Hugo Bastos[1], Francisco Couto[1],

[1] Department of Informatics, Faculty of Sciences,
University of Lisbon, Edifício C6, Piso 3
Campo Grande, 1749-016 Lisbon
{cmachado, hbastos}@xldb.di.fc.ul.pt, fcouto@di.fc.ul.pt

Abstract. Our understanding of biological systems is highly dependent on the study of the mechanisms that regulate genetic expression. In this paper we present a tool to evaluate scientific papers that potentially describe *Saccharomyces cerevisiae* gene regulations, following the identification of transcription factors in abstracts using text mining techniques. GREAT evaluates the probability of a given gene-transcription factor pair corresponding to a gene regulation based on data retrieved from public biological databases.

Keywords: Gene Ontology, gene regulations, *Saccharomyces cerevisiae*, text mining

1 Introduction

Cellular processes are regulated by interactions between various types of molecules such as proteins, DNA, RNA and metabolites. Among these, the interactions between transcription factors and their target genes play a prominent role, controlling the activity of proteins and the expression levels of genes.

The knowledge acquired in Molecular Biology can be retrieved essentially from two types of sources: scientific literature and public biological databases. In order to automatically identify gene regulations from both these sources, two approaches have to be devised: development of text-mining tools to extract specific biological entities [1], namely gene and transcription factor names; identification of biological databases containing gene annotations and other relevant information, such as experimental evidence.

This paper describes GREAT, a tool for the automated evaluation of scientific papers previously identified by a text mining tool as potentially describing *Saccharomyces cerevisiae* gene regulations. GREAT will be used together with the text mining tool to identify scientific papers and the gene regulations they reference. These will then be evaluated by Yeastract [2] curators and stored in the Yeastract database.

2 DataSources

Yeastract. The regulations contained in Yeastract are catalogued either as documented or potential: documented when the regulation was assessed either through methods that analyze the binding of the transcription factor to the target gene promoter region or that analyze changes in the target gene expression in consequence to the transcription factor suppression; potential when the only evidence found was the transcription factor binding motif in the target gene promoter region.

Entrez: PubMed and Gene Databases. Entrez is a retrieval system designed for searching linked databases that includes PubMed and Gene [3]. Abstracts can be obtained from the PubMed database and lists of genes referred in the abstract or in the whole article can be obtained from PubMed and/or Gene.

Gene Ontology and Gene Ontology Annotation (GOA). GO was born due to the need to describe and conceptualize biological entities in an unambiguous way. It provides a vocabulary describing the roles of genes and gene products in a species independent fashion, and is organized in three aspects: Molecular Function, Biological Process and Cellular Component [4].

The GOA project [5] is housed by the European Bioinformatics Institute and aims to provide high-quality GO annotations to proteins in the UniProt Knowledgebase (UniProtKB) and International Protein Index (IPI). Since the purpose of GO is to provide the vocabulary and not the actual annotation of the genes and genes products, GOA is used as a more complete source of annotations.

3 Identification of Transcription Factors in Scientific Literature

The software for this step is being developed in Python and comprises four modules responsible for the following tasks: obtention and storage of abstracts; identification of transcription factors in abstracts; identification and score attribution to selected text features (used to build a statistical model); classification of the abstracts as relevant or non-relevant for the purpose of gene regulations, using libbow's implementation of Support Vector Machines (SVM) [6].

4 Gene Regulation EvAluation Tool

For each abstract identified by the text mining tool as having at least one transcription factor, GREAT retrieves the genes referenced in it (or in the article) and calculates the probability of each gene-transcription factor pair corresponding to a gene regulation. The genes are obtained from the Entrez databases and the probability is calculated from the output of the steps described below and shown in Fig.1, through the use of a simple machine learning methodology such as decision trees [7].

Physical Potential. The identification of the Physical Potential consists in evaluating if the gene's promoter region contains the transcription factor binding motif. This is assessed using the list of potential regulations catalogued in Yeastract database: if a given pair is found in this list, we have a confirmation that the transcription factor can bind to the gene, and thus has the potential to control its transcription. This step has a binary output: existence/not existence of physical potential.

Biological Potential. The quantification of the Biological Potential consists in comparing the biological processes in which the gene and transcription factor are involved. This is done by means of a semantic similarity measure such as those described in [8] and [9], using the Biological Process aspect of GO. If the gene and transcription factor are involved in the same biological processes then it is likely that there is a regulation relationship between them. The output in this step is numeric, as defined by the semantic similarity measure used. The evaluation of the most suited measure is in progress.

Type of Evidence. The Type of Evidence refers to the experimental assay performed to obtain the information described in the article. Direct methods, such as Chromatin Immunoprecipitation assays (ChIP), provide more reliable results than indirect methods, such as microarrays. The higher the reliability of the method with which the gene-transcription factor pair was obtained, the higher the probability of a true regulation. Considering a list of direct and indirect methods obtained from Yeastract, a survey of public biological databases is being done in order to identify those that can provide this type of information.

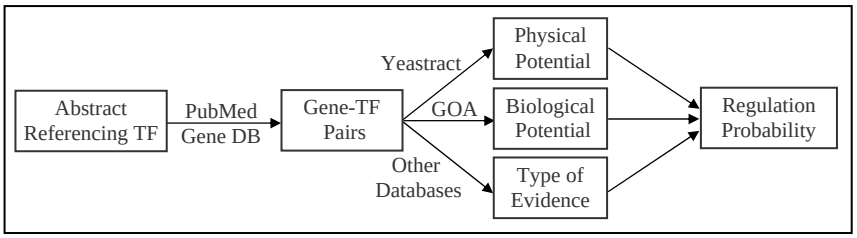

Fig. 1. GREAT framework: for each abstract referencing at least one transcription factor (TF), a list of all the genes it references is retrieved from PubMed and/or Gene databases; then the putative regulation between the transcription factors and each gene is analyzed in terms of Physical Potential (using Yeastract), Biological Potential (using GOA) and Type of Evidence (using other databases), in order to calculate the regulation probability.

Text Analysis. This task, which is a future work perspective, will consist in the extraction of text fragments from the abstract that may contain evidence that the gene-transcription factor pair corresponds to a true regulation.

5 Conclusions

The identification of the articles potentially describing gene regulations is in a more advanced stage than GREAT, although both are still in early stages of their development.

Currently, GREAT's gene retrieval step is fully implemented and the subsequent steps are in final stages of planning or in early stages of development. As a consequence, no quantitative results are available at the moment.

The selection of the databases for the Type of Evidence evaluation is proving to be the most difficult step, since not all databases annotate the PubMed identification number to their entries, and when they do, it is hardly for every entry.

As previously stated, the tools developed are intended to help the Yeastract curators in the identification of new *Saccharomyces cerevisiae* gene regulations in literature. In addition to this, the designed framework allows the automatic validation of Yeastract's potential regulations.

After the implementation of these tools depending only on the abstract, we plan to move on into the full text. Although present problems will be intensified, and new ones will emerge, we expect the articles' selection to be more accurate, and the validation steps will be eased, as the identification of the type of evidence.

Acknowledgments. This work was supported by FCT, through the project PTDC/EIA/67722/2006, the Multiannual Funding Program.

References

1. Rebholz-Schuhmmann, D., Kirsch, H., Couto, F.: Facts From Text - Is Text Mining Ready to Deliver? PLoS Biol 3: e65 (2005)
2. Teixeira, M.C. *et al*: The YEASTRACT Database: a Tool for the Analysis of Transcription Regulatory Associations in *Saccharomyces cerevisiae*. Nucl. Acids Res. 34, D446--D451 (2006)
3 Sayers, E.W. *e tal*: Database Resources of the National Center For Biotechnology Information. Nucl. Acids Res. 37, D5--D15 (2009)
4. Ashburner, M. *et al*: Gene Ontology: Tool for the Unification of Biology. Nature Genetics 25, 25--29 (2000)
5. Barrel, D. *et al*: The GOA Database in 2009 - An Integrated Gene Ontology Annotation Resource. Nucl. Acids Res. 37, D396--D403 (2009)
6. McCallum, A.K.: Bow: A Toolkit For Statistical Language Modelling, Text Retrieval, Classification and Clustering. http://www.cs.cmu.edu/~mccallum/bow (1996)
7. Breiman, L., Friedman, J, Stone, C.J., Olshen, R.A.: Classification and Regression Trees. Chapman & Hall, USA (1984)
8. Lord, P., Stevens, R., Brass, A., Goble, C.: Investigating Semantic Similarity Measures Across the Gene Ontology: the Relationship Between Sequence and Annotation. Bioinformatics 19, 1275--1283 (2003)
9. Pesquita, C. *et al*: Metrics for GO Based Protein Semantic Similarity: a Systematic Evaluation. BMC Bioinformatics Suppl 5, S4 (2008)

Changes made to the paper

Review 1:

Comments for the authors:
It would be appropriate to have presented some result

At the moment there are no quantitative results available, so it wasn't possible to satisfy the reviewer requirement. A sentence was added to the Conclusions section stating this.

Review 2:

Comments for the authors:
Major comments:

1. The article lacks enough coverage on the defined framework.

2. It will be very useful to state plausible solutions to the potential challenges mentioned in the article. e.g. what are the possible ways to combine the different potentials?

In the context of both these requirements, several actions were taken in section 4 - Gene Regulation EvAluation Tool: information was added, sentences were rewritten for clarity improvement, Fig. 1 and respective legend suffered alterations, also for clarity improvement.

3. State-of-the art in the area of "entity recognition in biological literature" is in a matured stage and should be covered in the article.

This comment led to our understanding that the paper wasn't sufficiently focused in GREAT. The text mining tool to identify articles potentially containing transcription factors and gene regulations is outside the scope of GREAT and this paper. We added a reference in the introduction to a review article regarding entity recognition in biological literature, but also made alterations to the abstract and introduction to clarify the scope of the paper.

Minor comments:

1. The article mentions "public databases" several times. It should be made clear that it is in reference to biological databases.

This requirement was satisfied throughout the paper.