RETRIEVING RELEVANT DOCUMENTS FROM BIOLITERATURE FOR THE CURATION OF YEAST GENE REGULATORY NETWORKS

H. P. Bastos[†1], C. Machado[1], A. T. Freitas[2], F. M. Couto[1]

[1] *Universidade de Lisboa, Portugal*
[2] *INESC-ID, Portugal*

[†] E-mail: *hbastos@xldb.di.fc.ul.pt*

ARN (Algorithms for the identification of genetic Regulatory Networks) aims at the development of methods that will partially automate the study, identification and modeling of mechanisms found in many living organisms that control gene expression. We are developing information integration tools for automatically identifying gene regulations from BioLiterature, since the vast growth of Bioliterature makes the manual curation process unfeasible. Techniques that exploit Bioliterature can perform information retrieval tasks, which classify and obtain relevant documents form a large corpus, or go further into information extraction tasks, in which relations between biological entities are automatically identified.
We performed a preliminary study of our approach with a subset of the documents that were the main source of information for building YEASTRACT, a curated repository of regulatory associations between transcription factors and target genes in Saccharomyces cerevisiae. Our study involved the classification of each phrase transcription factor occurrence using a machine learning approach based on support vector machines (SVM). The transcription factors occurrences were only taken from the abstracts of each document. We obtained a 88% precision with 89% recall, which show the effectiveness of our approach even at an early stage and encourage us to enhance our methods to achieve a even better performance. After completion our tool will be integrated in the curation process of the YEASTRACT database making it more efficient.
In the future, Gene Ontology will also be used as a confirmation of the biological significance of the retrieved regulatory data.