

BRINGING TEXT MINERS AND BIOLOGISTS CLOSER TOGETHER

Anália Lourenço¹, Sónia Carneiro¹, Rafael Carreira²,
Miguel Rocha², Isabel Rocha¹, Eugénio C. Ferreira¹

¹ IBB - Institute for Biotechnology and Bioengineering, Center of Biological Engineering

² Department of Informatics / CCTC

University of Minho

Campus de Gualtar, 4710-057 Braga – PORTUGAL

{analia,soniacarneiro, ecferreira,irocha}@deb.uminho.pt,
{rafaelcc,mrocha}@di.uminho.pt

Abstract

The boosting of Biomedical Text Mining (BioTM) research in the last few years has led the way for finally bridging out the gap between text miners and biologists. Beyond the development of enhanced entity recognisers and the construction of relationship extraction systems, now, more than ever, it is the time for applying available tools to real-world scenarios. Moreover, it is crucial to develop end-user tools that can assist biologists in their research activities. Such tools should be able to emulate biologist conventional curation, recurring to the same knowledge bases and making the same assumptions that biologists usually do, whereas delivering automated capabilities. The search and selection of PubMed articles, the construction of dictionaries from the contents of available Molecular Biology repositories, the implementation of description environments for rule specification, the implementation of dictionary- and rule-based entity recognisers, the development of flexible and extensible relationship extraction systems and the development of easy-to-use manual curation environments are of foremost importance.

Our software, named @Note, aims to be a framework and a workbench for BioTM, i.e., it has been conceived for delivering end-user applications, whereas enabling collaboration with other BioTM groups. As a framework, it provides a reusable design for BioTM software systems and a set of pre-assembled software building blocks that programmers can use, extend and customise for their specific needs. As a workbench, it helps developing BioTM applications by integrating Natural Language Processing and Data Mining tools and supporting major Information Retrieval and Information Extraction processes. Moreover, it encompasses a flexible and extensible manual curation environment that enables the interaction with biologists, correcting former annotations and enhancing dictionary contents. We successfully applied @Note in the study of the stringent response on *Escherichia coli*, an important subject within the analysis of stress responses in bacteria. This joint effort allowed biologists to contribute to the enhancement of our manual curation environment and to identify new functionalities for the existing plug-ins and the specification of new plug-ins.