Uncovering mechanisms underlying the *Escherichia coli* stringent response: application of text mining approaches to identify meaningful relations in text documents

Sónia Carneiro[1], Anália Lourenço[1], Rafael Carreira[2], Miguel Rocha[2], Eugénio C. Ferreira[1], Isabel Rocha[1]

[1]IBB - Institute for Biotechnology and Bioengineering, Centre of Biological Engineering
[2]Departament of Informatics / CCTC
University of Minho, Campus de Gualtar, 4710-057 Braga – PORTUGAL

Physiological stress responses in complex organisms are regulated by elaborated interactions. However, many of those interactions are not well-known and, in particular, most of the entities involved in the response have remained unidentified. The study of microbial stress responses in model-organisms like *Escherichia coli* (*E. coli*) is expected to expand the knowledge about stress response systems. During extreme environmental conditions, bacteria respond to changes by altering the gene expression pattern and protein activity. Therefore, the main scientific challenge here is the identification of the mechanisms that control the response to stress.

In bacteria, changes in the genetic expression during stress conditions are commonly controlled at the transcriptional level by regulatory proteins, transcription and sigma factors. While transcription factors bind to specific sequences of DNA promoting or blocking the activity of the RNA polymerase, sigma factors bind to the RNA polymerase reprogramming its ability to recognize different promoter sequences and express new sets of target genes. These complex biological systems encouraged us to retrieve the highest amount of information from literature to understand a particular stress response called stringent response[1]. This stress response, caused by amino acid starvation, has remained a topic of interest during several years due to its role in growth in organisms. So far, this response has been characterized by the accumulation of an unusual guanosine nucleotide, (p)ppGpp. Two proteins, RelA and SpoT, help to maintain the level of (p)ppGpp in the cell. However, the regulatory mechanisms behind these events are still unclear. Hypotheses are continuously emerging and the collection of all the information available is crucial in order to assist the investigation of the complex network behind the survival behaviour of bacteria under hostile conditions.

Text mining is being acknowledged by the potential to extract valuable information contained in scientific publications. The increasing interest in community-based annotation has led to the development of several tools capable of supporting the identification of important contents[2]. Combining mined text with other sources of evidence from biological databases, which may include a wide range of entities, such as genes, proteins and chemicals, can lead to the capture of novel biological evidences. In general, the search for important information involving a particular biological problem, such as a physiological response in an organism, can be widely supported by the implementation of these information extraction processes.

Our work addresses the study of the mechanism of stringent response in *E. coli*, combining a dictionary- and rule-based entity recognition system. From the collection of documents potentially related to *E. coli* stringent response that can be extracted from PubMed, the system allowed to extract all biological entities potentially involved that will be further analysed and that can be used to reconstruct a network model of the process. In order to validate the methodology used, we analysed the most frequently mentioned biological entities. From that analysis, we found that the system recognized the guanosine nucleotide (ppGpp), RelA and SpoT proteins and the transfer and ribosome ribonucleic acids (tRNA and rRNA, respectively) as the most frequent stated biological entities. Indeed, these entities have a central role in the stringent response mechanism, as mentioned previously. However, although entities like 50S ribosomal subunit protein L11, Fis transcriptional dual regulator and ribosome modulation factor (RMF) are not always reported as key players, they were found frequently mentioned in stringent response-related literature. In fact, they are involved in some processes acknowledged to be involved in stress responses. For example, the 50S ribosomal subunit protein L11 plays an important role in regulating the stringent response, since the ppGpp synthesis by RelA requires both uncharged tRNA at the A site of the ribosome and the presence of L11 protein[3]. The RMF has been associated with the negative translational control by facilitating the formation of inactive ribosome dimers under stringent circumstances[4]. Likewise, the DNA-bending Fis protein, required for efficient initiation of chromosome replication and for activation of stable RNA (rRNA and tRNA) genes, is also subject to stringent control[5].

Therefore, it is reasonable to conclude that text mining processes allows to collect in a systematic way entities address the discovery of facts and information of major importance to acknowledge biological relationships, minimising both resource and labour efforts and accounting for the ever-expanding literature.

1. L. U. Magnusson, A. Farewell, T. Nystrom, *Trends in Microbiology* 2005**,** 13 (5), 236-242.

2. K. B. Cohen, L. Hunter, *PLOS Computational Biology* 2008**,** 4 (1).

3. T. M. Wendrich, G. Blaha, D. N. Wilson, M. A. Marahiel, K. H. Nierhaus, *Molecular Cell* 2002**,** 10 (4), 779-788.

4. K. Izutsu, A. Wada, C. Wada, *Genes to Cells* 2001**,** 6 (8), 665-676.

5. O. Ninnemann, C. Koch, R. Kahmann, *Embo Journal* 1992**,** 11 (3), 1075-1083.