BioSys-PhD
Biological Systems - Functional & Integrative Genomics

Project N - Supervisor: Francisco Couto (LASIGE) | Co-supervisor: Luka Clarke (BioFIG)

Title: Development of a Text Mining Approach to Disease Network Discovery**.**

Objectives: Creation of a text mining-based framework for elucidation of biological networks in disease, using transcription factor/target interactions as a case study**.**

Methodology:

An important effort in any area of research is the organization of knowledge in meaningful and systematic ways; this ensures that the community can easily search for state-of-the-art information on relevant research. A fundamental question in molecular biology is how transcription factors act in the context of perturbed networks of gene expression in disease. Current work in the supervisor's group extending the power of text mining methodologies will be combined with ongoing work in the co-supervisor's laboratory on transcriptional networks in health and disease, specifically cystic fibrosis, to develop a framework for elucidation of the most likely transcription network hubs, based on published literature. Existing resources, such as TransmiR (http://www.cuilab.cn/transmir), will be mined to determine how expression of key transcription factors is altered in diseased tissues, and how this is reflected in the altered expression of transcription factor targets (in this case miRNAs). Output from the framework will be useful in drug target discovery in the study of human disease.

The goal of the proposed framework will be to compare and cluster scientific texts in order to categorize them in meaningful groups. The framework will be designed to deal with several biomedical domains, however its main focus will be information relevant for understanding transcription factors. The framework will initially implement text mining techniques for recognizing transcription factors and their relationships to diseases from biomedical literature and ontological knowledge as domain knowledge to guide and validate the results (see http://dx.doi.org/10.1371/journal.pone.0062984). The text mining techniques will test and explore two approaches: (i) a rule-based approach that relies on rules inferred from patterns identified from the text by experts. The rules represent in a structured form the knowledge acquired by experts when performing the same task. The expert analyzes a subpart of the text and identifies common patterns in which the relevant information is expressed. These patterns are then converted to rules to identify the relevant information in the rest of the text. (ii) a case-based approach that relies on a predefined set of texts previously annotated by an expert, which is used to learn a model for the rest of the text. Cases contain knowledge in an unprocessed form, and they only describe the output expected by the users for a limited set of examples. The expert analyzes a subpart of the text (training set) and provides the output expected to be returned by the text-mining system for that text. All the collected information will be stored, organized and integrated according to semantic web techniques (see http://dx.doi.org/10.1093/bib/bbt079), and its exploration will test and explore enrichment analysis techniques (see http://dx.doi.org/10.1093/bib/bbt079), and semantic similarity measures (see http://dx.doi.org/10.1142/S0219720013710017) that have already been successfully applied to other biomedical domains.

Supervisor: Francisco M.Couto **Co-Supervisor:** Luka Clarke

Type of fellowship:
National (Portugal)