Protein-centered biological networks by automatic caption analysis

Enrico M. Bucci

BioDigitalValley Srl, Italy

Abstract

In former years, a lot of attention has been paid to the retrieval of meaningful biological information connecting proteins and genes, i.e. relationships between different players in the cascade of molecular events regulating the physiology and pathology of cells, tissues and eventually organisms. The main goal is to develop genes/proteins connection models able to explain complex biological phenomena in terms of emerging properties of large, structured networks, whose topology and detailed structure account at least in part for these properties. This implies the use of experimental methods able to collect information on a large number of different proteins under different conditions, and then properly connecting the data to the results obtained all over the world, so to get a coherent picture in a larger frame. In particular, to encompass a larger body of information and to figure out how some experimental study fits to the accumulated knowledge, methods are required to retrieve the available data on all proteins involved in the study (the target proteins), as well as on all proteins, which are connected by some piece of information to the targets. To this aim, a method consists in parsing automatically the scientific literature, retrieving co-occurring names of proteins, genes or other kinds of molecules and attempting to identify some terms which qualifies the relationship between the identified proteins. This task is a non trivial one, giving the ambiguity in gene/protein nomenclature (which affects both precision and recall of the relevant data), and the strong dependence of the type of relationship on the context at multiple levels. Most of the available methods parse only the abstracts of the scientific literature; however, the information contained in the abstracts is often incomplete, due to the fact that only those genes/proteins which are in the main scope of the paper are discussed, while often data on a number of other proteins are contained elsewhere.

In an attempt to overcome these limitations, we focussed on the analysis of the figure captions contained in the scientific literature. The captions of a paper refer in most cases to the experiments described in the paper, and thus contain an enriched amounts of data describing the biology of different proteins, including the relationship between them. Moreover, since terms referring to gene/proteins and other terms related to experimental methodologies are simultaneously present in a reduced textual space, it is possible to identify groups of proteins studied with a certain experimental technique; by properly filtering for a specific technique, is possible to characterize the type of relationship between the proteins. For example, proteins co-occurring in a caption describing

a double-hybrid experiment are most likely binding partners, while proteins co-occurring in a caption describing a 2D-gel experiments are probably co-expressed in a given condition/biological sample.

We thus developed Protein Quest, a tool which automatically and efficiently parse both the abstract and the captions of scientific paper in a pdf document. Results obtained from more than 2.000.000 free, full-text papers will be discussed, with reference to the topological characterization of the obtained cooccurrence networks and to the dependence of their topology from different query strategies; moreover, some specific, disease-oriented networks and predictions will be presented.