Bibliome Informatics from a Complex Networks Perspective

Luis Rocha
Indiana University, United States
Instituto Gulbenkian de Ciencia, Portugal, rocha@indiana.edu

KeyWords: *Literature mining; Text Mining; Bibliome; Complex Networks; Information Extraction; Knowledge organization;*

Much of the research presently conducted in the biomedical domain relies on the induction of correlations and interactions from data. Because we ultimately want to increase our knowledge of the biochemical and functional roles of genes and proteins in organisms, there is a clear need to integrate the associations and interactions among biological entities that have been reported and accumulate in the literature and databases. Biomedical literature mining is an important informatics methodology for large-scale information extraction from repositories of textual documents, as well as for integrating information available in various domain-specific databases and ontologies, ultimately leading to knowledge discovery, vizualization, and organization. It helps us tap into the biomedical collective knowledge, and uncover relationships and interactions buried in the literature and databases, and even those inferred from global information but unreported in individual experiments. Our approach to literature mining is based on bottom-up, complex network, data-driven or bio-inspired methods, which we have applied to automatic discovery, classification, vizualization and annotation of protein-protein and drug-drug interactions, pharmacokinetic data, protein sequence family and structure prediction, functional annotation of transcription data, enzyme annotation publications, and so on. In this talk we will overview some of our results, focusing on the mapping of collective knowledge via weighted networks of large-scale co-occurrence data. A few of our publications on the topic are listed below:

[1] Wu, Hengyi, et al [2013]. *An Integrated Pharmacokinetics Ontology and Corpus for Text Mining.* BMC Bioinformatics. BMC Bioinformatics. 14:35. (Highly Accessed)

[2] A. Kolchinsky, A. LourenÃğo, L. Li, L.M. Rocha [2013]. *Evaluation of linear classifiers on articles containing pharmacokinetic evidence of drug-drug interactions.* Pacific Symposium on Biocomputing, 2013. 18:409-420.

[3] A. LourenÃğo, M. Conover, A. Wong, A. Nematzadeh, F. Pan, H. Shatkay, and L.M. Rocha [2011].*A Linear Classifier Based on Entity Recognition Tools and a Statistical Approach to Method Extraction in the Protein-Protein Interaction Literature.* BMC Bioinformatics. 12(Suppl 8):S12

[4] A. Kolchinsky, A. Abi-Haidar, J. Kaur, A.A. Hamed and L.M. Rocha [2010]. *Classification of protein-protein interaction full-text documents using text and citation network features.* IEEE/ACM Transactions On Computational Biology And Bioinformatics, 7(3):400-411.

[5] A. Abi-Haidar, J. Kaur, A. Maguitman, P. Radivojac, A. Retchsteiner, K. Verspoor, Z. Wang, and L.M. Rocha [2008]. *Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks.* Genome Biology. 9(Suppl 2):S11

[6] Maguitman, A. G., Rechtsteiner, A., Verspoor, K., Strauss, C.E., Rocha, L.M. [2006]. *Large-Scale Testing Of Bibliome Informatics Using Pfam Protein Families.* In: Pacific Symposium on Biocomputing 11:76-87.

[7] Verspoor, K., J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L.M. Rocha, T. Simas [2005]. *Protein Annotation as Term Categorization in the Gene Ontology using Word Proximity Networks.* BMC Bioinformatics, 6(Suppl 1):S20.