Towards integrating large-scale OMICS data to better understand disease pathogenesis

Mohamed Hamed
Institute for Biostatistics and Informatics in Medicine and Ageing Research (IBIMA), University of Rostock Medical Center, Rostock, Germany

Abstract—The biological functions of the molecular elements of genes, transcription factors, and noncoding RNAs as well as mutations thereof are tightly linked to cell development, malfunctions, and disorder pathways. Moreover, those molecular elements interact with each other forming complex regulatory network that governs, on one hand, regular cellular pathways, and on the other hand, their dysregulation or malfunction in pathological processes. Therefore, uncovering this regulatory architecture in complex living systems is being considered as one of the most critical challenges of modern systems biology.

Nowadays, advances in sequencing and expression technologies enable the generation of large high-throughput datasets that allow for genome-wide association studies and facilitate scrutinizing the regulation mechanisms between various molecular elements. In the light of the availability of genomic, transcriptomic, and epigenomic data from different protocols and experiments, new integrative approaches are needed to integrate such Omics data into a coherent description, from which researchers may derive new well-founded hypotheses. Consequently, this would hint at boosting the probability of identifying genetic key players and critical regulatory pathways that could drive complex diseases and tumorigenesis.

Herein, we present some of our published bioinformatics approaches developed as open source software tools to integrate heterogeneous sources of large-scale Omics data and uncover the combinatorial regulatory interactions between different genetic elements. First, we present a general integrative network-based approach that involves transcriptional and post--transcriptional interactions and reports the computational analysis of gene and miRNA transcriptomes, DNA methylome, and somatic mutations. This workflow enables users to identify putative disease drivers and novel targets for therapeutic treatment.

For linking the somatic mutations with other genomic data sets, a stand-alone pipeline named "SnvDMiR" was implemented to investigate possible genomic proximity relationships between somatic variants and both differentially methylated CpG sites as well as differentially expressed miRNAs.

Next, we developed TFmiR as a freely available web server for deep and integrative downstream analysis of combinatorial regulatory interactions between TFs/genes and miRNAs that are involved in the etiology of human diseases.

We demonstrated the efficacy of our approaches by identifying biologically relevant network modules and potential driver genes/miRNAs/mutations in breast cancer, diabetes, and Parkinson's disease (PD). Moreover, we utilized our approaches to investigate the molecular mechanisms of cellular differentiation (namely hematopoiesis) as an example for biological processes.

In summary, our integrative approaches have the potential to provide biologically relevant analysis as demonstrated on breast cancer data and PD data. The presented approaches are also independent on the different experimental assays of expression profiling. It can be applied in a similar fashion to study cell development and other diseases progression. The provided topological and functional analyses of our approaches promote them as reliable bioinformatics tools for researchers across the life science communities.

Keywords—Systems biology; integrative Omics; Gene regulatory network; high-throughput data analysis