CeTICSdb: an integrated platform for analysis of heterogeneous, high-throughput -omics and mathematical modeling of biochemical reactions

Milton Y. Nishiya-Ya Jr1, Marcello S. Reis1, Daniel F. Silva1,2, Eduardo S. Kitano1, Glauzia M. Souza1, Inácio L.M. Junqueira de Azevedo1, Julia P.C. da Cunha1, Junior Barreto-Câmara1, Leo K. Iwai1, Solange M.T. Serrano1, Hugo A. Armelin2

1 LETA-CeTICS, Instituto Butantan, Brazil; 2 Programa Unidisciplinar em Bioinformática, Universidade de São Paulo, Brazil; 3 Escola Politécnica, Universidade de São Paulo, Brazil; 4 Instituto de Matemática e Estatística, Departamento de Ciência da Computação, Universidade de São Paulo, Brazil; 5 Instituto de Química, Departamento de Bioquímica, Universidade de São Paulo, Brazil.

e-mail: milton.nishiya@butantan.gov.br

Introduction

The Center of Toxins Immune-response and Cell Signaling (CeTICS) is an emerging center that studies biochemical, molecular, and cellular mechanisms of toxins that have therapeutic potential, aiming to understand the short-term and long-term behavior of biological systems based on analyses of signaling networks. Once those studies and analyses involve collaborations across disciplines such as Biology, Mathematics, Statistics, and Computer Science, the research developed in CeTICS is intrinsically interdisciplinary. This fact, coupled to the heterogeneous, high-throughput data produced by modern high scale techniques in genomics and proteomics, implies the necessity of data organization and integration to carry out scientific knowledge. To this end, we are developing CeTICSdb, a platform to provide a uniform conceptual schema to minimize -omics data representation, as well as a suitable semantics for interdisciplinary research that allows quantitative and qualitative -omics integrative analyses and mathematical modeling of signaling networks.

Main Features of the CeTICSdb platform

The CeTICSdb accommodates the features described in the Introduction. Moreover, it was designed to allow the integration of the different -omics data and the mathematical modeling of kinetic models, which in turn includes enzymatic signaling networks. The platform is composed of modules, which will allow to efficiently involve it into a data management framework, requiring far fewer manual changes, especially in the development of new applications.

The database has been developed in a conceptual modeling for data integration, combining data from different sources and formats. The database will store semantics of biological experiments, few and processed high-throughput data, such as genome and transcriptome sequences, proteome peptides, gene and protein expression profiling, besides annotation and ontologies.

Analysis of high-throughput Transcriptome of Snakes

Proteome Statistical Analysis

In a qualitative high-throughput proteomics analysis, the main goal is the study of protein expression profiling (such as differentially expressed proteins, enrichment of protein sets, principal component analysis and clustering methods), using computational and statistical methodologies.

To assess the metabolic pathway (KEGG) level patterns in observed expression changes, to understand the interaction within and between pathways, and identify sets of proteins co-regulated or correlated activated/deactivated pathways, we developed a tool, to calculate the pathway Activity Scores (AS), enabling statistical correlation and hierarchical clustering analysis between the pathways in different conditions (Fig 3), and a protein set enrichment analysis (Fig 4) based on Gene Ontology database.

These analyses were applied to the quantitative proteome (label free) of sugarcane leaves, from two ancestral species (Saccharum officinarum [Soff] and S. spontaneum [Ssp]) and a hybrid cultivar (SPB-D160 [SPb]), that differ from each other in their capacity to produce sucrose and to resist to pathogens, to study their expression and interaction between pathways from different cultivars and to identify the most enriched GO terms, given a set of proteins.

Conclusions and Perspectives

The CeTICSdb platform manages and stores RNA-Seq projects and their respective data, starting from the main stages of the processing pipeline. For instance, the workflow of the analysis of high-throughput transcriptomic data of the Bothrops jaranaca snake includes:

- Extraction of multiplexed samples based on: single end, paired-end, mate-pair.
- Assembly of high-quality reads for transcriptome without reference.
- Mapping against reference (or de novo transcriptome assembly) and to databases.
- Estimation of transcript abundance (RPKM/FPKM) based on edgeR, DESeq, etc.

Differentially Expressed genes in RNA-Seq with correlated expression profile are clustered (Fig.5) and the transcript clusters, extracted. Hierarchically clustered Spearman correlation matrix resulting from comparing the transcript expression values for each pair of samples (Fig.6).

The CeTICSdb platform in its present conceptual database model allow construction and simulation of dynamic models based on high-throughput omics data. In addition, it also allows storage, management, retrieval and integration of experimental and simulated results.

However, further development in CeTICSdb database structure and data integration tools is planned to reduce gap between behavior of experimental models and virtual simulated models.