Published online: 2019-08-16

Appendix: Content Summaries of Selected Best Papers for the 2019 IMIA Yearbook, Section Bioinformatics and Translational Informatics

Lee SI, Celik S, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, Estey EH, Miller CP, Chien S, Dai J, Saxena A, Blau CA, Becker PS

A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia

Nat Commun 2018 Jan;9(1):42

The authors present in this paper a statistical method to identify molecular markers for targeted treatment of acute myeloid leukemia using omics data (genome-wide gene expression profiles) and in vitro sensitivity to 160 chemotherapy drugs. They describe the MERGE algorithm (standing for mutation, expression hubs, known regulators, genomic copy number variation, and methylation) a computational method to identify gene expression markers using multi-omics data. In a nutshell, MERGE learns from data the contribution of five key features (*e.g.*, mutation associated to acute myeloid leukemia, hubness in a gene expression network) to the drive of gene potentially implicated in cancer progression. A complete approach is designed ranging from data collection and method development to both *in silico* and *in vivo* validation.

Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, Brat DJ, Cooper LAD

Predicting cancer outcomes from histology and genomics using convolutional networks

Proc Natl Acad Sci U S A 2018;115(13):E2970-E2979

This paper presents a method to predict the survival of patients based on digital pathology images as well as genomics biomarkers. The authors developed a neural network based approach to predict the survival of patient on the basis of digitalized pathology images and genomics biomarkers. The authors describe a Survival Convolutional Neural Network (SCNN) designed to predict the survival of patients suffering from glioma. The networks are trained using public data coming from the TCGA datasets. To help in the interpretation and understanding of the prediction, the authors use a heat map visualization highlighting the structures identified as important by the neural networks.

Sengupta S, Sun SQ, Huang KL, Oh C, Bailey MH, Varghese R, Wyczalkowski MA, Ning J, Tripathi P, Mc Michael JF, Johnson KJ, Kandoth C, Welch J, Ma C, Wendl MC, Payne SH, Fenyö D, Townsend RR, Dipersio JF, Chen F, Ding L

Integrative omics analyses broaden treatment targets in human cancer

Genome Med 2018 Jul 27;10(1):60

In this work the authors adopt a pan-cancer approach to take benefit of multi-omics data for drug repurposing. Their goal is to identify drugs approved by the Food and Drug Administration for cancer location not yet mentioned in the approval. The authors rely on the Database of Evidence for Precision Oncology (DEPO), a tool they built, to link druggability to genomic, transcriptomic, and proteomic biomarkers. They used a pan-cancer cohort of more than 6,500 tumors to identify tumor with potential druggable markers. The authors rely on the DEPO database (integrating genomic, transcriptomic, proteomic data, and clinical data over several types of cancer samples) and structural alignment tools for identifying tumors with potentially druggable biomarkers consisting of drug-associated mutations, micro-RNA expression outliers, and protein/phosphoprotein expression outliers. Orthogonal validation of putative biomarkers was performed thanks to the large-scale drug screening dataset GDSC (Genomics of Drug Sensitivity in Cancer).

Torshizi AD, Petzold LR

Graph-based semi-supervised learning with genomic data integration using condition-responsive genes applied to phenotype classification

J Am Med Inform Assoc 2018;25(1):99-108

This paper presents a graph-based semi-supervised method to phenotype classification of samples. Several graphs of labeled and unlabeled samples are built on features sets corresponding to distinct genomic levels (gene expression, DNA methylation, micro-RNA). Additional graphs add pathway knowledge -for each considered genomic level- through the use of condition-responsive genes (CORGs). CORGs are, for each pathway, the most discriminative set of genes containing the highest statistical signal level. The authors define three feature sets corresponding to different subsets of CORGs (the whole sets, the top P-value ranked genes, the top ranked genes according to their frequency in all CORGs). A weighted integration of the various graphs is performed before the semi-supervised learning based on the K nearest neighbors principles. The method was applied on ovarian cancer data from the Human genome Atlas. The conducted experiments assessed the added value of the method compared to the existing ones. The results also show that the classification accuracy is effectively improved when integrating transcriptomic, epigenetic, and pathway knowledge.