Integrated Cancer Subtyping using Heterogeneous Genome-Scale Molecular Datasets

Suzan Arslanturk† and Sorin Draghici
Department of Computer Science, Wayne State University
Detroit, MI 48202, USA
Email: suzan.arslanturk@wayne.edu
sorin@wayne.edu

Tin Nguyen
Department of Computer Science and Engineering, University of Nevada
Reno, Nevada 89557, USA
Email: tinn@unr.edu

Vast repositories of heterogeneous data from existing sources present unique opportunities. Taken individually, each of the datasets offers solutions to important domain and source-specific questions. Collectively, they represent complementary views of related data entities with an aggregate information value often well exceeding the sum of its parts. Integration of heterogeneous data is therefore paramount to i) obtain a more unified picture and comprehensive view of the relations, ii) achieve more robust results, iii) improve the accuracy and integrity, and iv) illuminate the complex interactions among data features. In this paper, we have proposed a data integration methodology to identify subtypes of cancer using multiple data types (mRNA, methylation, microRNA and somatic variants) and different data scales that come from different platforms (microarray, sequencing, etc.). The Cancer Genome Atlas (TCGA) dataset is used to build the data integration and cancer subtyping framework. The proposed data integration and disease subtyping approach accurately identifies novel subgroups of patients with significantly different survival profiles. With current availability of vast genomics, and variant data for cancer, the proposed data integration system will better differentiate cancer and patient subtypes for risk and outcome prediction and targeted treatment planning without additional cost and precious lost time.

Keywords: data integration, disease subtype discovery, omics data

1. Introduction

Genomic and epidemiologic studies over the past decade have generated a wealth of data, including molecular, variant, and clinical data on both individuals and populations that can be leveraged to better understand cancer risk, progression, and outcomes. Subtyping patient disease populations using high-dimensional molecular data has transformed how researchers and clinicians interpret and quantify heterogeneity within a disease. Subtyping has been highly effective in discovering cancer types, tumor histologies, survival rates, treatment planning and responses. Investigation of clinically relevant disease subtypes cannot be achieved by using a single dataset in isolation from others due

Work partially supported by grants NIH/NIDDK (1R01DK107666-01), Department of Defense (W81XWH-16-1-0516), and National Science Foundation (SBIR 1853207)