Chapter 105
Personalized Disease Phenotypes from Massive OMICs Data

Hans Binder
University of Leipzig, Germany

Kathrin Lembcke
University of Leipzig, Germany

Lydia Hopp
University of Leipzig, Germany

Henry Wirth
University of Leipzig, Germany

ABSTRACT

Application of new high-throughput technologies in molecular medicine collects massive data for hundreds to thousands of persons in large cohort studies by characterizing the phenotype of each individual on a personalized basis. The chapter aims at increasing our understanding of disease genesis and progression and to improve diagnosis and treatment. New methods are needed to handle such "big data." Machine learning enables one to recognize and to visualize complex data patterns and to make decisions potentially relevant for diagnosis and treatment. The authors address these tasks by applying the method of self-organizing maps and present worked examples from different disease entities of the colon ranging from inflammation to cancer.

INTRODUCTION

Application of new high-throughput technologies in molecular medicine such as microarrays and next generation sequencing generates massive amounts of data for each individual patient studied. These methods enable to characterize the genotype and/or molecular phenotype on a personalized basis with the aim to increase our understanding of disease genesis and progression and, in final consequence, to improve diagnosis and treatment options. New methods are needed to handle such 'big data' sets collected for hundreds to thousands of persons in large epidemiological cohort studies, e.g. to accomplish data mining and classification tasks with impact for diagnosis and therapy. From the perspective of bioinformatics and systems biomedicine, 'big data' challenge objectives such as data integration, di-

DOI: 10.4018/978-1-4666-9840-6.ch105

mension reduction, data compression and visual perception. To finally achieve a personalized therapy it is necessary to link genetic variations to molecular disease phenotypes, to associate molecular with clinical data, to extract, to filter and to interpret bio-medical information and finally, to translate these discoveries into medical practice.

Machine learning represents one interesting option to tackle these tasks. Particularly, neural network algorithms such as self-organizing maps (SOMs) combine effective data processing and dimension reduction with strong visualization capabilities. These methods provide a suited basis to analyze large and complex data generated by modern bioanalytics.

The present contribution shortly describes the method of 'SOM portraying'. We demonstrate data compression capabilities which reduce the dimension of the relevant (in terms of functional information) data by several orders of magnitude. The strong visualization capabilities of the SOM approach are illustrated. They enable the comprehensive, intuitive and detailed analysis of 'big data' in molecular medicine by mapping them into phenotype and feature space. To illustrate the performance of the method we present a series of representative case studies from different disease entities and OMICs realms related to the human colon.

BACKGROUND

Big Data from High-Throughput Bioanalytics

Standard medical practice is moving from relatively ad-hoc and subjective decision making to so-called evidence-based healthcare which makes use of complex diagnosis technologies such as comprehensive laboratory analyses and powerful imaging techniques.

Powered by the progress in modern molecular biomedicine the number and granularity of accepted disease types and also the variety of related therapy options steeply increase. This trend is paralleled by increasing volume and complexity of data collected per patient in disease-related cohort studies and also in medical practice. Accordingly, the way of decision making in diagnosis changes, for example from evaluating a set of key laboratory markers to information mining in large and potentially 'big' data sets generated by high-throughput technologies. Moreover, the evaluation of currently collected data includes also their comparison with already accumulated knowledge and reference data which itself can constitute a 'big' data challenge.

As generally accepted, big data is characterized by the three (Beyer, 2011), and sometimes four 'V': big volume, big velocity, big variety and, also, big veracity referring usually to the scale of the data, the handling of streaming data, the manifold and complexity of different forms and values of data and to their uncertainty, respectively. For high-throughput data in molecular medicine these general criteria can be specified: Usually the number of single data items per sample measurement ranges from tens of thousands to several millions and even more depending on the type of data (e.g. proteomics measured by means of mass spectrometry or genomics measured by means of next generation sequencing) and on their level in the processing pipeline starting with raw data and ending with highly (information-) enriched data (see below). In this respect present- and next generation omics-technologies generate massive amounts of data. Velocity in terms of time needed to store and re-store the data and to process