Analysis of Large-Scale OMIC Data Using Self Organizing Maps

Hans Binder
Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany

Henry Wirth
Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany

INTRODUCTION

The development and decaying costs for high-throughput bio-molecular analytics give rise to huge and still increasing amounts of data collected in the context of modern 'omics' realms. These studies aim at discovering the functioning of life on different molecular level as subsumed by the 'omes' such as the genome, transcriptome or proteome. The experimental data generated require the design of adequate and powerful analysis strategies and methods. Tasks such as transformation of measured data into calibrated features, their appropriate evaluation and weighting according to their importance in the biological context and suited support for extraction and interpretation of information become extremely puzzling tasks. Machine learning using neural network algorithms represents one interesting option to tackle them.

The information processing capabilities of the human brain are highly effective and reached by no means by modern computers in many aspects. It appears desirable to make use of the potential of neuronal perception, abstraction and decision making and to apply such 'natural' principles in 'artificial' computer algorithms. The method of 'self-organizing maps' (SOM) applies concepts of neuronal data perception to the processing of vast amounts of information. It occurs as a promising attempt to analyze molecular-biological high-throughput data because it accomplishes essential tasks such as clustering, dimension reduction, multi-dimensional scaling and visualization.

Machine learning and particularly SOM are still somewhat unorthodox methods in life and health sciences. In consequence application of the concept of SOM learning, data transformation and visualization still require special explanation and adaptation. Moreover, the SOM algorithm accomplishes 'only' basal sorting and visualization tasks. It needs to be supplemented with add-ons for significance testing and marker extraction, visualization of biological properties inherent in the data and finally for information mining of the biological context to become an attractive application tool in life sciences.

BACKGROUND

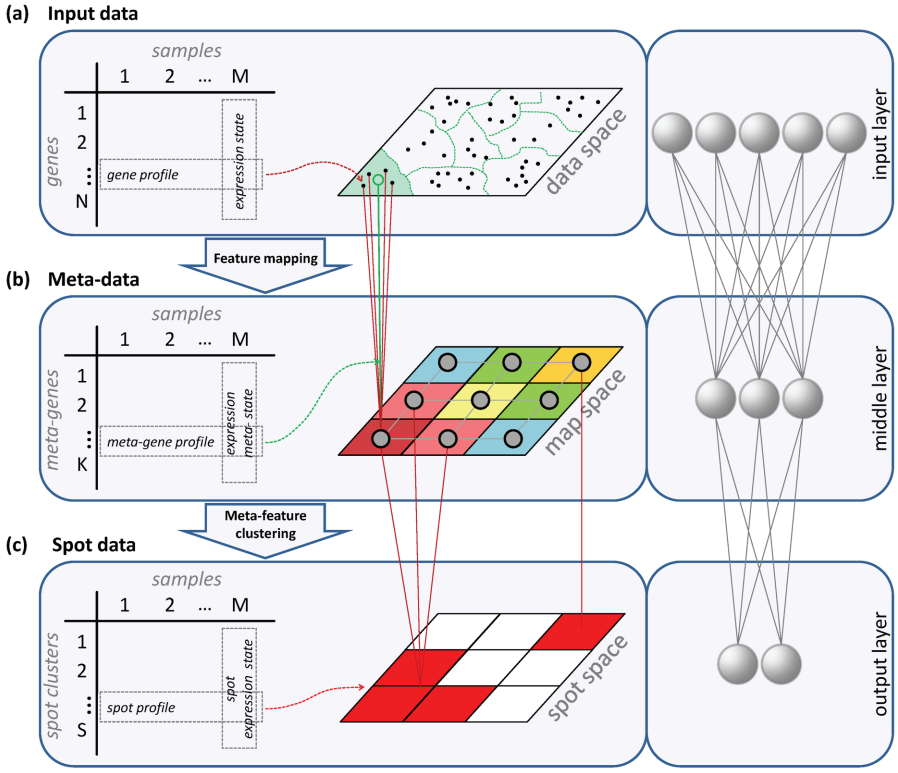
The SOM method was developed in the early 1980'ties by T. Kohonen (Kohonen, 1982). First applications to microarray gene expression data were published in 1999 (Tamayo et al., 1999) emphasizing a gene-centered perspective to cluster gene expression profiles into predefined groups of similarly expressed genes. A complementary sample-centered clustering approach was realized shortly after providing a visual identity of the expression landscapes of each sample (Golub et al., 1999). In the last years, SOM machine learning was also used to analyze proteomics and metabolomics data, and just first applications of self-organizing maps to epigenetics were published (Steiner et al., 2012).

THE SOM PORTRAYING METHOD

The data produced by high-throughput bioanalytics is usually given as a feature matrix of dimension N x M (see Figure 1) where N is the number of features measured per sample and M is the number of samples referring, e.g., to different treatments, time points or individuals. As a convention, each row of the matrix will be termed *profile* of the respective feature. The columns on the other hand will be termed *states* refer-

DOI: 10.4018/978-1-4666-5888-2.ch157

Figure 1. Two-step data compression using SOM machine learning: Firstly, the input data are transformed into meta-data where each meta-feature is trained such that its profile resembles that of a cluster of input features. Secondly, meta-data are clustered into 'spots' of similar meta-features. Data reduction topology of the SOM resembles that of Neuronal Nets as shown on the right.



ring to each of the conditions studied. In general, the number of features can range from several thousands to millions, depending on the experimental screening technique used. Typically, this number largely exceeds the number of states studied, i.e. N>>M. SOM machine learning aims at reducing the number of relevant features by grouping the input data into clusters of appropriate size, and thus to transform the matrix of input data into a matrix of so-called meta-data with a reduced number of meta-features, K<<N (Figure 1a and b). In other words, SOM aims at mapping the space of the high-dimensional input data onto meta-data space of reduced dimensionality.

The method to reach this aim is inspired by our assumptions about the perception of visual information in the brain. Accordingly, optical input stimuli are projected onto the neuronal net in the cortical area. Then, the connections between the neurons adapt to the visual pattern in a learning process. This causes self-organization of the neuronal network such that it better matches the activation pattern. The SOM approach mimics this input-driven self-organization where the 'stimuli' are given by the input data to which the meta-data adapt in an iterative learning process.

Learning

The learning step is in the heart of the SOM method. It starts with appropriate initialization of the map space, followed by the training process to adjust its intrinsic structure to the structure of the input data and ends with the final mapping and visualization of the map space in terms of SOM portraits, metagene profiles and different supporting maps.

Linear initialization effectively determines the initial values of the meta-features in all samples also called 'profiles' by utilizing the two-dimensional subspace spanned by the two largest principal components of the input data. Then, the SOM training algorithm itera-