Abstract

About 5% of all metastatic cancer cases fall into the category of cancer of unknown primary site (CUP). As the origin of cancer plays a critical role in choosing the optimal therapy for patients with metastatic cancer, there is an urgent need for a firm histologic diagnosis. Several studies propose gene expression profiling (GEP) for facilitating the identification of the site of origin of the primary tumour. Tothill *et al.* (Cancer Res. 2005. 65(10):4031–40) developed support vector machine (SVM) classifiers by using cDNA microarray data. In a previous project, the classifier was reconstructed in R using the publicly available microarray data. Comparable accuracy was achieved with the microarray data, but using the classifier on 21 Medical University Graz RNA-seq samples did not yield any meaningful results.
The overall aim of this project was to investigate the classification possibilities of RNA-seq CUP samples, firstly test how using raw microarray intensity data influences classification accuracy of SVM and neural network models and whether it improves the classification accuracy of RNA-seq data. The highest validation accuracy was achieved with SVMs trained on log_2-transformed intensity data (84.6%) and the lowest with raw intesity values (69.2%)
Further three different types of neural networks were realized and trained on TCGA RNA-seq data (feed forward neural network [FFNN], deep learning FFNN and a convolutional neural network [CNN]). While adjusting network shapes and parameter two approaches for classification have shown promising results. A sinlge hidden layer FFNN and a CNN yielded 97.4% and 98.5% classification accuracy respectively, while deep learning DLNN yielded 79.7% classification accuracy.
The realized neural network and SVM models were then used to classify 21 CUP samples provided by the MU Graz. This led to varying results depending on the methods used (highest accuracy: 52.38%).