



PREDICTING PRIMARY ORIGIN OF CANCER FROM SAMPLES MUTATION PROFILES

Hana Susak¹,

¹Centre for Genomic Regulation (CRG)

Abstract:

When cancer is found in one or more metastatic sites but the primary site cannot be determined, it is called a cancer of unknown primary origin (CUP). The American Cancer Society estimates that about 33,770 cases of cancer of unknown primary origin will be diagnosed in 2017 in the United States. This number represents about 2% of all cancers with usually poor prognosis. We introduced a classification model for predicting cancer primary origin using publicly available (exome) sequencing data for 3357 samples diagnosed with one of 6 different cancer types. To be more precise, we used Linear discriminant analysis (LDA) on sparse Partial Least Squares (sPLS) components to avoid the common problem in genomics - high number of features with much smaller number of samples. Only by using genomic features from sequencing data we achieved the accuracy of 0.65, compared to primitive classifier (majority class) of 0.29 accuracy. With more features (e.g. samples gender and age), we expect that accuracy can further be improved. We hope that accuracy can be improved to an even higher level, so that clinicians can use our model as a guide and help when choosing the most appropriate treatment for patients diagnosed with CUP.

Keywords:

Cancer prediction, Cancer of Unknown Primary, Classification, Partial Least Squares regression.

Correspondence: Hana Susak

e-mail: hana.susak@crg.eu