Method to predict the percentage of cell types in human blood

Anna Igolkina, Maria Samsonova

St.Petersburg Polytechnical University, Polytechnicheskaya 29, igolkinaanna11@gmail.com

Motivation and Aim:

Blood is the most investigated heterogeneous tissue. It contains a variety of cell types, of which the major types are Lymphocytes, Monocytes, Granulocytes, Erythrocytes, Megakaryocyte (Lymphocytes and Granulocytes are complex cell groups in turn). Gene expression data from blood genomics studies is widely used in medical diagnosis. Most of these studies are based on the analysis of total peripheral blood mononuclear cells (PBMCs). PBMCs are composed of over a dozen cell types, the proportion of which varies in blood samples from individual people. This variability significantly influences genome-wide gene expression data. The heterogeneity of blood distorts the data, however, it is often discarded due to the lack of data on the composition of the samples. The application of experimental methods to separate or quantify constituents from each sample is time-consuming and does not solve the problem. Therefore, an attractive alternative is to accurately deconvolve gene expression data. Here we develop a method to predict the percentage of cell types in a blood sample from whole genome gene expression data.

Materials:

We check our approach on two independent studies that we arrange by combing the available data from databases. The first study contained mouse gene expression samples obtained as mixtures of liver, brain and kidney with known proportions and pure cell-type samples. Mixture samples were bisected into test and training sets pure cell-type samples were defined as validation set. In the second study we worked with 4 human blood datasets. The largest of them contains 2000 patients with known gene expression levels and percentages of 5 cell types in blood samples. The samples were divided into training set (300 samples), testing set and set for prediction. Validation in the second study was apply to remain 3 datasets with pure cell type samples and whole blood samples (mixtures of 5 blood cell-types).

Methods and Algorithms:

We built and tested various predictive models based on PCA, linear regression model (with and without prior knowledge of cell type specific signatures obtained from pure cell types [1-2]) and SVM with different kernel types and two level linear regression approach. The last method showed the best predictive ability. To select a gene subset which provides the best prediction of cell proportions we construct heuristic feature selection algorithm consisted of censoring, filtration by objective function, and consistent subsampling. Prediction on genes obtained by this feature selection procedure showed better results than prediction on specific marker genes for blood cell-types[3].

It is noteworthy that both feature selection and predictive methods were constructed for each individual cell type independently. To estimate the performance of different approaches a the Pearson correlation coefficient between estimated and true cell type proportion in data was calculated.

Results:

We achieved the best estimation of cell type proportions using our heuristic feature selection procedure, and two level linear regression approach. This approach significantly improved the Pearson correlation between true and estimated cell type proportions to approximately 0,8-0.95 in both studies (mouse and human samples). This result is high enough for further studies.

Conclusion:

We have developed a method that can accurately predict the percentage of cell types from whole genome gene expression data in mouse artificial samples and human blood samples. Our approach can be used to predict the percentage of cell types in other tissues.

Availability:

The MATLAB script is available on request from the author.

The text of the abstract: up to four pages, Times New Roman, 12 pt, 1.5 interval.

Acknowledgements:

We thank Lude Franke (UMCG, Groningen) for provision with main human blood dataset of our work, Alexandra Zhernakova(UMCG, Groningen) for comments that improved biological understanding of the problem and Natalia Kadyrova (St.Petersburg Polytechnical University) for the great consultation in SVM theory.

1. Alexander R Abbas et al. (2009) Deconvolution of blood microarray data identifies cellular activation patterns insystemic lupus erythematosus, PloS one 4.7.

2. Ting Gong et al. (2011) Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples, PloS one 6.11.

3. Renaud Gaujoux (2013) An introduction to gene expression deconvolution and the CellMix package