Volume 50(3-4):153, 2006
Acta Biologica Szegediensis
http://www.sci.u-szeged.hu/ABS

DISSERTATION SUMMARY

Advances in gene expression based molecular diagnosis

János-Zsigmond Kelemen

Institute of Plant Biology, Biological Research Center, Hungarian Academy of Sciences, Szeged, Hungary

Global transcription profiling with DNA microarray technology has lead to a deeper understanding of the sophisticated cellular processes. Pathological alteration, as a complex biological process, is constantly being studied in this manner in a quest to find key drug targets. However, the large data sets comprising simultaneous expression levels of thousands of genes monitored under diverse circumstances still constitute great challenge for biologists as well as computational algorithm developers. It is known that various treatment procedures may have different effects on patients diagnosed as having the same type of cancer due to different origins or courses in the development of the tumor. Although patients suffering from leukemia may have similar symptoms, it has been shown that microarray generated gene expression patterns are capable of making the distinction between the different subtypes of the disease (Golub et al. 1999). Over the last few years many molecular classification approaches based on statistics or machine learning algorithms have been applied to microarray data. Their common feature is that they try to model classes of *a priori* annotated samples by means of supervised training. With the obtained model parameters they predict the belonging of an un-annotated sample to one of the known classes. So far the support vector machine (SVM) has been shown to have the best performance for microarray classification problems. It has been successfully applied with a variety of binary and mutli-class tumor classifications (Ramaswamy et al. 2001). Notable performance was also obtained with artificial neural networks (ANN) (Khan et al. 2001).

Here we propose the use of the linear Kalman filter (Kalman 1960) as a preprocessing step in microarray based molecular diagnosis. Taking into account the expression covariance between genes is desired in such classification problems, since this stands for the functional relationships that govern tissue state. Hereby, we show that employing the Kalman state estimator to remove functional noise yields linearly separable data, suitable for most classification algorithms.

It is known that microarray data are usually corupted with measurement noise from various sources. Some percentage of the variance of a measured gene-expression signal is also due to biological variation. We sought to use the Kalman filter to remove measurement and functional noise, modeled as normally distribuited random variable and to estimate the biological state. Therefore we built a simple measurement state-space model:

$$x_i=x_{i-1}+w_i$$
$$y_i=x_i+v_i$$

where y_i is a numerical vector containing the expression values measured from the ith sample, x_i is the filtered expression data and thus the biological state, and v_i and w_i are noise and biological, allowed variation respectively. To reduce dimensionality, we applied singular value decomposition (Alter et al. 2000). We then employed Kalman filtering on the obtained model with tuning beeing done solely on the training dataset.

We applied the method on publicly available datasets and we found that it boosts the performance of the widely used ANN, SVM, k-nearest neighbours and classification trees, improving diagnosis accuracy. Kalman filtering also greatly improves the graphical viasualization of microarray data.

References

Alter O, Brown PO, Bolstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 97(18):10101-10106.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531-537.

Kalman RE (1960) A new approach to linear filtering and prediction problems. J Basic Engineering 35-45.

Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 7:673–679.

Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR (2001) Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA 98(26):15149-15154.

Supervisor: László Puskás
E-mail: kelli@brc.hu