Maximum significance clustering of oligonucleotide microarrays

Dick de Ridder, Marcel J.T. Reinders
Information & Communication Theory Group
Faculty of Electrical Engineering, Mathematics & Computer Science
Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands
{D.deRidder,M.J.T.Reinders}@ewi.tudelft.nl

Frank J.T. Staal, Jacques J.M. van Dongen
Department of Immunology, Erasmus MC, University Medical Center
Dr. Molewaterplein 50, 3015 GE Rotterdam, The Netherlands
{F.Staal,J.J.M.vanDongen}@erasmusmc.nl

Abstract

Affymetrix high-density oligonucleotide microarrays measure expression of DNA transcripts using probesets, i.e. multiple probes per transcript. Usually, these multiple measurements are transformed into a single probeset expression level before data analysis proceeds; any information on variability is lost. In this work we demonstrate how individual probe measurements can be used in a statistic for differential expression. Furthermore, we show how this statistic can serve as a clustering criterion. A novel clustering algorithm using this maximum significance criterion is demonstrated to be more efficient with the measured data than competing techniques for dealing with repeated measurements, especially when the sample size is small.

1. A linear model

Suppose there are K conditions in a microarray study. Let the number of arrays measured under each condition be A_k, $k = 1, \ldots, K$ and $A = \sum_{k=1}^{K} A_k$. Each probeset g ($g = 1, ..., G$) is represented on each array by P_g probes. The following ANOVA model can be applied to the normalised, $\log_2$-transformed intensity measured for probe p on array a under condition k:

$$\log_2(X_a^{p,k}(g)) = \mu_g + \alpha_{g,p} + \beta_{g,k} + \varepsilon_{g,p,k,a} \qquad (1)$$

The expression level of probeset g under condition k is then $e_{g,k} = \mu_g + \beta_{g,k}$. An F-ratio can be used to assess the significance level at which the null hypothesis of no condition effect, $H_0(g) : e_{g,1} = e_{g,2} = \ldots = e_{g,K}$, can be rejected. Under H_0, this ratio is Fisher distributed with $K - 1$ df for the numerator and $P_g(A - K)$ df for the denominator.

2. Maximum significance clustering

In calculating the F-ratio, it is assumed it is known which samples belong to a condition k. However, in clustering, the goal is exactly to find this membership. To this end we can introduce an $A \times K$ membership matrix $\mathbf{Q}$, in which $Q_{ak} = 1$ if array a belongs to cluster k, and 0 otherwise. This allows us to write $(\mathbf{X}_{\cdot,\cdot}^{\cdot,k})\mathbf{1}$ as $\mathbf{XQ}$. For notational purposes, it is more convenient to use a slightly different membership matrix $\mathbf{M}$, with $M_{ak} = 1/\sqrt{A_k}$ if array a belongs to cluster k, and 0 otherwise: $\mathbf{M} = \mathbf{Q}(\mathbf{Q}^T\mathbf{Q})^{-\frac{1}{2}}$. The F-ratio can then be written as (with $\mathbf{M} > \mathbf{0}$, $\mathbf{M1} = \mathbf{1}$):

$$F(\mathbf{X}(g), \mathbf{M}) = (A - K)\frac{\mathbf{1}^T\mathbf{XMM}^T\mathbf{X1}}{\text{tr}\,(\mathbf{X}(\mathbf{I} - \mathbf{MM}^T)\mathbf{X}^T)} \qquad (2)$$

Maximising (2) w.r.t. $\mathbf{M}$, subject to the constraints on $\mathbf{M}$, will find an assignment of the arrays to clusters, such that the difference in expression between clusters is maximally significant. However, as the number of probes (and hence the df of the Fisher distribution under H_0) may differ between probesets, this cannot easily be extended to assigning cluster labels based on the data of multiple genes. Assuming independence between probeset expression levels, we therefore minimise the log of the combined p-values instead:

$$\log\left[p(H_0|\mathbf{X}, \mathbf{M})\right] = \sum_{g=1}^{G} \log\left[p(H_0|\mathbf{X}(g), \mathbf{M}) + r\right] \qquad (3)$$

with r an arbitrarily small non-zero regularisation factor; in all experiments here, $r = 10^{-300}$.

In [4], a hill-climbing algorithm is proposed maximising (2) w.r.t. a crisp membership matrix $\mathbf{M}$. This algorithm is called MSCK, for maximum significance K-clustering.

Although we are not aware of any previous model-based approaches to clustering oligonucleotide microarray data based on probe data, there is literature on the use of repeated measurements in clustering in general. In [7] an error-weighted clustering method is proposed using weights based on different measures of variability (denoted D^0, D^s and D^c here); and two possible distance measures based on the Kullback-Leibler divergence between two distributions are given in [5] (denoted D^{Jeffreys} and D^{Resistor}).

3. Experiments

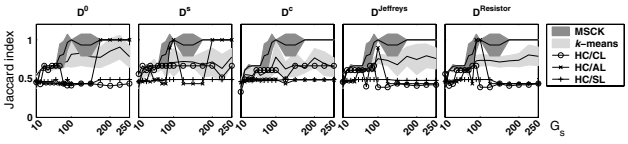
We applied the MSCK algorithm on both simulated data and some real-world data sets, and compared it to a number of standard clustering algorithm – k-means and hierarchical clustering with complete, average and single linkage – using the five distance measures D outlined above.

Our simulations (data not shown, see [4]) showed that MSCK performs well for small sample sizes A, seemingly combining the advantage of k-means (working well for small A) with that of the hierarchical methods (working well for larger K). It is most useful when the number of differentially expressed probesets and their expression difference is relatively small. For clearer differences between conditions, using variability information is not necessary.

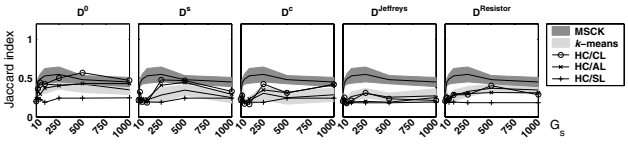
The real-world datasets used were: (a) a small subset of the Yeoh precursor B-ALL dataset [6], containing 16 BCR-ABL and 21 MLL samples, measured by HG-U95Av2 microarrays ($A = 37$, $K = 2$); (b) Pomeroy dataset A [3], containing 42 cases of five types of central nervous system embryonal tumor, measured on Hu6800 microarrays ($A = 42$, $K = 5$) and (c) a dataset of 7 development stages of T-cells [1] measured on two HG-U133A microarrays each ($A = 14$, $K = 7$)[1]. For all datasets, array background was first removed, arrays were quantile normalised [2] and the G_s probesets with most variation in probeset expression level over the A arrays were selected for use in clustering ("variation filtering").

Figure 1 shows the results as Jaccard indices between the known labels and the cluster assignments (mean $\pm$ sd over 10 random initialisations; 1 indicating perfect agreement, 0 none), as a function of G_s. On all datasets, MSCK performs reasonably well, especially for small G_s, showing it to be more efficient with the data than traditional methods. For larger G_s performance decreases, as the method starts to fit noise. Although for each dataset a clustering algorithm/distance measure combination can be found for which performance is comparable to that of MSCK, no single combination consistently outperforms it. On the Pomeroy and T-cell data, MSCK gives the best results; in fact, only MSCK is able to perfectly cluster the T-cell data.

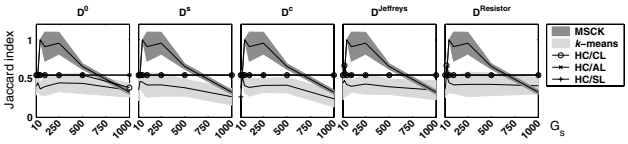
[1]For the HG-U133A microarrays in the T-cell dataset, which carry less probes per probeset, we found that a few outlier probes could have a very large influence. We therefore pre-processed this data for all methods by removing 3 outlier probesets, i.e. those with the highest average absolute deviation from average probe rank over all arrays.



(a) Yeoh data, BCR-ABL vs MLL: $A_1 = 16$, $A_2 = 21$



(b) Pomeroy dataset A: $A_1 = 8$, $A_2 = 4$, $A_3 = A_4 = A_5 = 10$



(c) T-cell development data: $A_i = 2$, $K = 7$

Figure 1. Experimental results (see text).

References

[1] W.A. Dik et al. New insights into early human T-cell development based on combined quantitative t-cell receptor gene rearrangement studies and gene expression profiling. *Journal of Experimental Medicine*, 2005. Accepted for publication.

[2] R.A. Irizarry et al. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):e15, 2003.

[3] S.L. Pomeroy et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442, 2002.

[4] D. de Ridder, M.J.T. Reinders, F.J.T. Staal, and J.J.M. van Dongen. Maximum significance clustering of oligonucleotide microarrays. Technical Report ICT-2005-03, Information & Communication Theory Group, Delft University of Technology, 2005. `http://ict.ewi.tudelft.nl/`.

[5] E. Wit and J. McClure. *Statistics for microarrays - design, analysis and inference*. John Wiley & Sons, Chichester, UK, 2004.

[6] E.J. Yeoh et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1(2):133–143, 2002.

[7] K.Y. Yeung, M. Medvedovic, and R.E. Bumgarner. Clustering gene-expression data with repeated measurements. *Genome Biology*, 4(5):R34, 2003.