



Quality Assessment of Microarray Data and Optimal Filtering Criteria

Takashi Kido¹, Janos Demeter², and Gavin Sherlock¹

Departments of Genetics¹ and ²Biochemistry
Stanford University School of Medicine
Stanford, CA 94305 USA



The Problem

What's the problem ?

The quality assessment of microarray data is an important issue, to prevent the risk of analysis of poor quality data. However, the methods for standard quality assessment have not yet been established. In addition, researchers use *ad hoc* filtering criteria to select data for analysis, because there is little or no guidance on optimal filtering parameters.

Motivation

- Generate quality metrics for individual microarrays, and microarray datasets.
- Learn how best to apply quality filters to microarray data using these metrics.

Background

Microarray data quality control is difficult due to the diversity of instrumental and biological factors. Labor intensive manual inspection of hundreds or thousands spots is still a common procedure. The field of microarray quality control has been largely neglected in the past but has recently become an area of interest.

Recent work includes the following topics:

- Automatic evaluation of microarray spot quality
- Semi automatic classification of spot quality
- Assignment of the quality weights to each microarray
- Imputation methods of the missing data
- Visualization of regional biases



However, these studies did not provide enough evidence on which quality metrics are better and how best to apply quality filters. Little systematic work has been done on quantitative evaluation of the quality assessment of microarray data and filtering criteria.

Quality Assessment Measure: Q-Score

Quality Assessment by "Q-Score"

The Q-Score is a quality measure for an array based on a subset of the spots present on the array. It is calculated with different fractions of the data removed, based on a spot metric's value, and can be used to guide selection of a cut-off value for that filter to remove low quality spots.

Definition of Q-Score

It is very common that the expression of a gene is measured multiple times, using different sequence probes (reporters), on the same array. The average spread of the log-ratios of reporters mapping to the same gene may be used to assess the quality of an array. The Q-Score is a measure of the 'within gene' spread of the data, and can be calculated using the formula on the Figure 1. A lower value of this score means a narrower spread of the data and a better quality array, while a higher value a wider spread of the data and lower quality array.

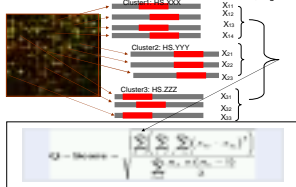


Figure 1: Q-Score formula. nk: number of reporters in a gene; m: number of genes with more than 1 reporter; x: Log ratio measurement for feature i of gene k minus the mean log-ratio for gene k.

Optimal Filtering Criteria

Is there any benefit to combining filters ?

Figure 4. Q-Score Curves on Different filtering criteria

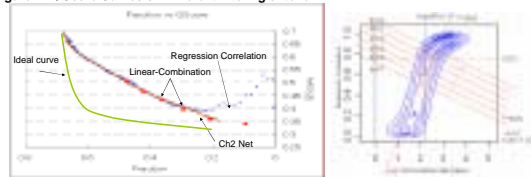
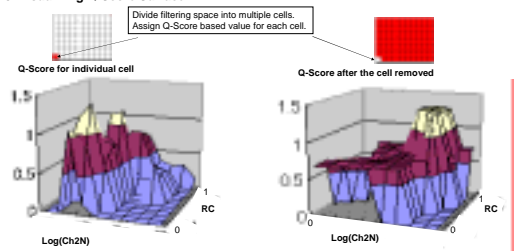


Figure 4 shows an example of Q-Score filtering dynamics when combining two different filters. In this case, linear combination of two filters slightly improves the Q-Score filtering dynamics. We are now investigating effective filtering criteria with multiple filters. The goal is to find effective filtering criteria which dramatically improve the Q-Score dynamics, such as shown in 'ideal curve' in the graph.

What are the best filtering criteria ?

Figure 5. Visualizing Q-Score Surface



Visualizing the Q-Score Surface might help us to find the best filtering criteria. Figure 5 shows the example of such a visualization. In this example, we divide the filtering space into cells in a two dimensional grid, and calculate the Q-Score either for each cell (left hand side) or the entire grid excepting a given cell (right hand side). In the left 3D graph, the summit part of the Q-Score surface identifies the low quality spots on the array, while on the right hand side, the summit identifies the high quality spots on the array.

Developing Software Tool

We have been developing a quality assessment and filtering tool for the two-channel microarray platform based on the Q-Score dynamics. This tool provides functions to compare the Q-Score dynamics for various filtering criteria. This GUI tool is implemented using Perl-Tk, and can run on the PC and Macintosh platforms.

Menu widget

The current version of this application provides the following functions for helping to find the optimal filtering criteria.

Load Image Data

This widget loads micro array raw data and image data.

Filtering Dynamics Analysis

This widget generates the Q-Score vs. Fraction graph for the specified filters. Then it detects the inflection points of the filtering curves and also calculates the filter rankings based on the filtering performance.

Visualizing Q-Surface

This widget visualizes the Q-Score surface described in the previous section. The user can examine the relationships between Q-Score and two filters.

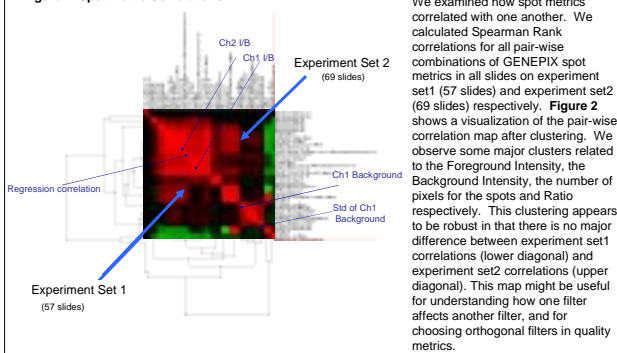
Visualizing Spot Images

This widget displays the spot images with filter values close to the value at the inflection point.

Multiple Filters

Spot Metrics: How do they correlate with each other?

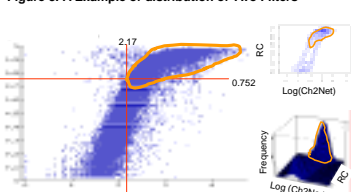
Figure 2. Spot Metric Correlations



We examined how spot metrics correlated with one another. We calculated Spearman Rank correlations for all pair-wise combinations of GENEPIX spot metrics in all slides on experiment set1 (57 slides) and experiment set2 (69 slides) respectively. Figure 2 shows a visualization of the pair-wise correlation map after clustering. We observe some major clusters related to the Foreground Intensity, the Background Intensity, the number of pixels for the spots and Ratio respectively. This clustering appears to be robust in that there is no major difference between experiment set1 correlations (lower diagonal) and experiment set2 correlations (upper diagonal). This map might be useful for understanding how one filter affects another filter, and for choosing orthogonal filters in quality metrics.

Combining filters ?

Figure 3. A Example of distribution of Two Filters



We are now investigating the effect of combination of filters. We might find the better filter and cutoff line by combining different filters. Figure 3 shows the example of distribution of two different filters, regression correlation and log(ch2 normalized Net). By applying multi-dimensional filtering, we might find the better cutoff surfaces rather than single cutoff point of one dimensional filtering.

Summary and Future Directions

Summary

1. Q-Score is reasonably consistent with visual spot evaluations.
2. Inflection points of Q-Score may be useful for determining the optimal cutoff for filtering data.
3. Spot metrics can be clustered. Correlation matrix among dozens of filters is very robust. This can be used to choose orthogonal filters.
4. Combining multiple filters may increase filtering efficiency. Finding the adaptive optimal filtering criteria is one of our research challenges.
5. We have been developing a GUI tool for quality assessment and filtering. Our goal is to further develop this tool, such that it can suggest optimal filtering parameters.

Future Directions

1. Examine the consistency with other quality metrics in the wide range of data sets. Generate overall quality metrics.
2. Develop the algorithm for adaptively finding the optimal filtering criteria with multiple filters.
3. Applying Machine learning for quantitative relationships between filters and Q-Score dynamics.
4. Further develop GUI tools for quality assessment and filtering.