Protocol

Calculation of Spot Reliability Evaluation Scores (SRED) for DNA Microarray Data

Kazuro Shimokawa, Rimantas Kodzius, Yonehiro Matsumura, and Yoshihide Hayashizaki

This protocol was adapted from “Methods for Increasing the Utility of Microarray Data,” Chapter 6, in *DNA Microarrays* (ed. Schena). Scion Publishing Ltd., Bloxham, UK, 2007.

INTRODUCTION

In terms of cost per measurement, the use of DNA microarrays for comprehensive and quantitative expression measurements is vastly superior to other methods such as Northern blotting or quantitative reverse transcriptase polymerase chain reaction (QRT-PCR). However, the output values of DNA microarrays are not always highly reliable or accurate compared with other techniques, and the output data sometimes consist of measurements of relative expression (treated sample vs. untreated) rather than absolute expression values as desired. In effect, some measurements from some laboratories do not represent absolute expression values (such as the number of transcripts) and as such are experimentally deficient. This protocol addresses one problem in some microarray data: the absence of accurate measurements. Spot reliability evaluation score for DNA microarrays (SRED) offers a reliability value for each spot in the microarray. SRED does not require an entire microarray to assess the reliability, but rather analyzes the reliability of individual spots of the microarray. The calculation of a reliability index can be used for different microarray systems, which facilitates the analysis of multiple microarray data sets from different experimental platforms.

RELATED INFORMATION

Superior microarray data produce superior results. Input the highest-quality microarray data possible, taking care to manufacture and hybridize the microarrays using the most rigorous scientific procedures. Poorly printed microarrays and low-quality samples produce inferior raw data, which will negatively affect the downstream computational processes. Superior robotics, printing technology, surface chemistry, target and probe preparation, and other molecular aspects produce superior data for analysis. Also, in advance of calculating SRED values, it is advantageous to obtain as many QRT-PCR control measurements as possible. If fewer than 50 control measurements are used and the data seem dubious, obtain additional QRT-PCR measurements. The parameter data for calculating SRED scores can be improved further by using improved algorithms and by adding new types of external data to the analyses. See also **Calculation of Absolute Expression Values for DNA Microarray Data** (this issue), which addresses the problem that some microarray data sets fail to provide absolute expression levels.

MATERIALS

CAUTIONS AND RECIPES: Please see Appendices for appropriate handling of materials marked with <I>, and recipes for reagents marked with <R>.

Equipment

Personal computer running Windows 2000 (or newer version) with at least an Intel Pentium IV CPU, 3.6 GHz processor, and 800 MHz front side bus

For convenience, statistical software packages such as R or SYSTAT (Systat Software Inc.) are recommended for use in calculating SRED scores.

METHOD

1. For each spot or gene element in the microarray, collect all of the parameter data listed in Table 1.
2. Prepare as many QRT-PCR control measurements corresponding to as many microarray spots as possible.
3. Place each microarray spot for which there is a corresponding QRT-PCR measurement into one of two groups, $R_n \leq 1$ or $R_n > 1$, using the following equation:

$$R_n(t_a, c_a, t_p, c_p) = \left| \frac{\left| \log_2(t_a/c_a) - \overline{\log_2(t_a/c_a)} \right|}{\sigma_a} - \frac{\left| \log_2(t_p/c_p) - \overline{\log_2(t_p/c_p)} \right|}{\sigma_p} \right|$$

where σ_a is the standard deviation of the microarray data, σ_p is the standard deviation of the QRT-PCR data, t_a is the intensity of microarray test data, c_a is the corresponding control value of the spot, t_p is the intensity of QRT-PCR sample, and c_p is the corresponding QRT-PCR control intensity.

4. Find the most suitable combination of parameters to explain the experimental results (i.e., which group a given microarray spot belongs to) by applying discriminant analysis.
5. Calculate SRED scores for all of the microarray spots using the parameters obtained in Step 4 and the following equations:

$$d_t^2(X) = (X - m_t)' S_p^{-1} (X - m_t)$$

$$D_t^2(X) = d_t^2(X) - 2 \ln(q_t)$$

$$p(t | X) = \exp\left(-\frac{D_t^2(X)}{2}\right) / \sum_u \exp\left(-\frac{D_u^2(X)}{2}\right)$$

where X is a vector of values corresponding to the parameters in Table 1, t is a subset of the group of test data for which the difference of the expression intensities of microarray and QRT-PCR is less than twofold, $d_t^2(X)$ is the squared Mahalanobis distance from X to group t , m_t is a vector containing variable means in group t , S_p is the pooled covariance matrix in group t , $D_t^2(X)$ is the generalized squared distance from X to group t , q_t is the prior probability of membership in group t , and $p(t | X)$ is the posterior probability of an observation X belonging to group t , $u = \{t, \bar{t}\}$.

DISCUSSION

Spot Reliability Evaluation

The abundance of small-scale microarray experiments using various experimental platforms has yielded large quantities of data and rendered database deposition of results an important process for the entire research community. Expression data are usually associated with supplementary data such as RNA sample information, experimental methods used, and so forth. When storing expression data in databases, it is important to store accompanying detailed experimental conditions. Access to such information and the subsequent integration of the microarray measurement data are necessary for detailed analysis of any microarray data set. The Microarray Gene Expression Data Society has defined a standard known as the Minimum Information About a Microarray Experiment (Ball et al. 2004). This

Table 1. Candidate parameters for discriminant analysis obtained from DNA microarray data

Parameter	Category	Description
1	Signal intensity of each microarray spot	Average of replicate signal intensities of channel 1 (e.g., Cy3) and channel 2 (e.g., Cy5)
2	Signal intensity of each microarray spot	Minimum signal intensity of channel 1 (e.g., Cy3) and channel 2 (e.g., Cy5)
3	Difference between the signal intensity and background intensity of each microarray spot	Average of the difference between the signal intensity and the background intensity of channel 1 (e.g., Cy3) and channel 2 (e.g., Cy5)
4	Difference between the signal intensity and background intensity of each microarray spot	Minimum difference between the signal intensity and the background intensity of channel 1 (e.g., Cy3) and channel 2 (e.g., Cy5)
5	Ratio of signal intensity to background intensity of each spot	Average of the ratio of the signal intensity to the background intensity of channel 1 (e.g., Cy3) and channel 2 (e.g., Cy5)
6	Ratio of signal intensity to background intensity of each spot	Minimum value of the ratio of the signal intensity to the background intensity of channel 1 (e.g., Cy3) and channel 2 (e.g., Cy5)
7		Reproducibility of the measurement value for each spot calculated as the difference between replicate experiments: $\log_2(t_{\text{array}1}/c_{\text{array}1}) - \log_2(t_{\text{array}2}/c_{\text{array}2})$, where $t_{\text{array}1}$ is the intensity of test data from microarray 1, $c_{\text{array}1}$ is the corresponding control value of the spot from microarray 1, $t_{\text{array}2}$ is the intensity of test data from microarray 2, and $c_{\text{array}2}$ is the corresponding control value of the spot from microarray 2
8		Relative expression value of each gene calculated as: $\log_2(t_{\text{array}}/c_{\text{array}})$
9		Overall reproducibility of the microarray (Pearson's correlation coefficient of duplicate experiments)

standard outlines the minimum information that should be reported with a microarray experiment to ensure its unambiguous interpretation and reproduction.

The number of publicly available data sets has increased dramatically in recent years and the need for tools and methods to compare and integrate these data sets has grown equally quickly. The SRED reliability index offers the end user an intelligible metric by which to estimate the accuracy of expression values for each spot in a DNA microarray. This method estimates the accuracy of expression data by applying multivariate analysis, using multiple data sources originating from the actual microarray spot measurements. The SRED score is calculated principally using multivariate discriminant analysis. The analysis parameters are chosen by comparing the difference between expression values derived from the DNA microarray experiment and corresponding measurements obtained by QRT-PCR; SRED is subsequently calculated for each spot (gene) in the DNA microarray not validated by QRT-PCR. An implicit assumption when calculating SRED scores is that QRT-PCR gives accurate reference expression values. For increased accuracy, primers for QRT-PCR can be selected from the RTPrimerDB public database (<http://medgen.ugent.be/rtprimerdb/>).

One way to represent spot intensity values from a DNA microarray is as the expression ratio of two samples such as $\log_2(\text{Cy3}/\text{Cy5})$, where the Cy3 and Cy5 values represent the measured intensities of each sample. SRED requires that the corresponding values of QRT-PCR be calculated in the same form. The difference, R , between microarray and QRT-PCR is given by the following equation:

$$R(t_a c_a t_p c_p) = |\log_2(t_a/c_a) - \log_2(t_p/c_p)|$$

where t_a is the intensity of microarray test data, c_a is the corresponding control value of the spot, t_p is the intensity of the QRT-PCR sample, and c_p is the corresponding QRT-PCR control intensity.

It is important to consider the variance in the measurements between different technologies. For instance, we have found previously that the standard deviation of expression ratios derived from QRT-PCR was 1.8 times larger than that of microarray experiments (Matsumura et al. 2005). It is therefore necessary to normalize these two values when calculating the difference. The normalized difference, R_n can be calculated using the equation described in Step 3. As an aside, we do not consider the non-linearity of the expression ratio between the techniques using this approach.

As mentioned above, SRED is calculated using multivariate discriminant analysis. The SRED score is defined as the probability $P(t)$ that the difference between the relative expression values of each spot (gene) determined using the DNA microarray or QRT-PCR is less than a factor of 2 (this condition is synonymous with $R_n \leq 1$). The reliability of spot intensities is evaluated using parameters obtained from the microarray experiment itself. We considered nine possible different parameters from the microarray experiment (summarized in Table 1). The most suitable combination of parameters for predicting the experimental result was decided by applying multivariate discriminant analysis to all of the possible combinations. In our case, a discriminant function using two parameters, 7 and 8 (see Table 1), proved to show the greatest efficiency and sensitivity when the threshold value was set to $R_n = 1$ (i.e., a twofold difference between microarray spot intensity and QRT-PCR expression value; Matsumura et al. 2005).

Theoretical Background of the Method

In the process of defining the SRED scores, we employed a combination of two approaches to improve predictive accuracy: Mahalanobis distances and Bayesian methods. In general, discriminant function analysis is used to determine variables that discriminate between two or more naturally occurring groups (usually called source groups). Conceptually, the distance between each sample and the center of every source group is computed in the multidimensional space described by the data properties (in this case, parameters 7 and 8 in Table 1), with the sample belonging to the closest source group. In this case, we used discriminant analysis to determine which source group ($R_n \leq 1$ or $R_n > 1$) a sample most likely belongs to. If Mahalanobis distances are used, the probability of the sample belonging to any source group can be calculated, because probabilities are inversely proportional to the Mahalanobis distances. In our case, Mahalanobis distances (a nonlinear distance model) perform better than the linear distance model.

In order to incorporate prior knowledge of the distribution of spots in the two groups, we use a Bayesian approach. Using the QRT-PCR and cDNA microarray test data, we can obtain a prior probability, $P(t)$, describing how likely a microarray spot is to be within twice the value of the corresponding QRT-PCR measurement. The term t denotes a subset of the group of the spot that satisfies $R_n \leq 1$, and $\bar{t}$ corresponds to $R_n > 1$. The prior probability enables us to calculate a posterior probability $\hat{P}(t | X)$: the probability that a spot that has the parameter vector X satisfies $R_n \leq 1$. The posterior probability will be the final SRED score (the spot reliability). For the final calculation of SRED scores, we combine both methods to obtain the equations used in Step 5:

$$\begin{aligned} d_i^2(X) &= (X - m_i)' S_i^{-1} (X - m_i) \\ D_i^2(X) &= d_i^2(X) - 2 \ln(q_i) \\ p(t | X) &= \exp\left(-\frac{D_t^2(X)}{2}\right) / \sum_u \exp\left(-\frac{D_u^2(X)}{2}\right) \end{aligned}$$

Both Mahalanobis distance and the Bayesian approach are well-known methods that have been used previously for multivariate discriminant analysis. In one implementation, SRED scores were assigned to approximately 1,500,000 spots in the Riken Expression Microarray Database (in this example, 133 QRT-PCR controls were used) (Bono et al. 2002). It is important to emphasize that the parameters used for the SRED calculation are specific to this microarray system. However, it is possible to apply SRED to other microarray systems by readjusting the parameters.

ACKNOWLEDGMENTS

We would like to thank Albin Sandelin and, for help with editing, Ann Karlsson. This work was supported (in part) by a grant from the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology, Japan. Rimantas Kodzius was supported courtesy of an FP5 INCO2 to JAPAN fellowship from the European Union.

REFERENCES

- Ball, C., Brazma, A., Causton, H., Chervitz, S., Edgar, R., Hingamp, P., Matese, J.C., Icahn, C., Parkinson, H., Quackenbush, J., et al. 2004. An open letter on microarray data from the MGED Society. *Microbiology* **150**: 3522–3524.
- Bono, H., Kasukawa, T., Hayashizaki, Y., and Okazaki, Y. 2002. READ: RIKEN Expression Array Database. *Nucleic Acids Res.* **30**: 211–213.
- Matsumura, Y., Shimokawa, K., Hayashizaki, Y., Ikeo, K., Tateno, Y., and Kawai, J. 2005. Development of a spot reliability evaluation score for DNA microarrays. *Gene* **350**: 149–160.

Cold Spring Harbor Protocols

Calculation of Spot Reliability Evaluation Scores (SRED) for DNA Microarray Data

Kazuro Shimokawa, Rimantas Kodzius, Yonehiro Matsumura and Yoshihide Hayashizaki

Cold Spring Harb Protoc; doi: 10.1101/pdb.prot4937

Email Alerting Service

Receive free email alerts when new articles cite this article - [click here](#).

Subject Categories

Browse articles on similar topics from *Cold Spring Harbor Protocols*.

[Analysis of Gene Expression](#) (197 articles)
[Analysis of Gene Expression, general](#) (129 articles)
[Analysis of Microarrays](#) (19 articles)
[Analysis of Microarrays, general](#) (13 articles)
[Bioinformatics/Genomics, general](#) (192 articles)
[Computational Biology](#) (100 articles)
[Expression Analysis of RNA](#) (14 articles)
[Microarrays](#) (54 articles)
[Microarrays, general](#) (55 articles)
[Molecular Biology, general](#) (1293 articles)

To subscribe to *Cold Spring Harbor Protocols* go to:
<http://cshprotocols.cshlp.org/subscriptions>