APPLICATION NOTE

Genomics & Informatics Vol. 5(3) 129-132, September 2007

arrayImpute: Software for Exploratory Analysis and Imputation of Missing Values for Microarray Data

Eun-Kyung Lee[1], Dankyu Yoon[2] and Taesung Park[3]*

[1]Department of Clinical Pharmacology and Therapeutics, University of Ulsan, Seoul 138-736, Korea, [2]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-921, Korea, [3]Department of Statistics, Seoul National University, Seoul 151-921, Korea

Abstract

arrayImpute is a software for exploratory analysis of missing data and imputation of missing values in microarray data. It also provides a comparative analysis of the imputed values obtained from various imputation methods. Thus, it allows the users to choose an appropriate imputation method for microarray data. It is built on R and provides a user-friendly graphical interface. Therefore, the users can easily use arrayImpute to explore, estimate missing data, and compare imputation methods for further analysis.

Keywords: imputation, microarray data, missing data

Introduction

Microarray experiments generate data sets for the expression levels of thousands of genes simultaneously. However, these experiments often produce missing values due to various reasons such as scratches on the chip, spotting problem, and the presence of dust. Most statistical methods for analyzing microarray data cannot be applied when the data contain missing values. Therefore, the missing values have to be estimated before further analysis of the microarray data.

Many imputation methods for the estimation of missing values have been developed, such as weighted k-nearest neighbors imputation (kNN, Troyanskaya *et al.*, 2002), Bayesian principal component analysis (BPCA, Oba *et al.*, 2003), local least squares imputation (LLS, Kim *et al.*, 2004), and robust least squares imputation with principal components (RLSP, Yoon *et al.*, 2006).

Even though many sophisticated imputation methods are currently available, the performanceof the proposed imputation method mainly depends on the characteristics of the missing data. Further, it has been shown that even a small number of poorly estimated missing values might produce misleading results (Wang *et al.*, 2006). Therefore, it is important to use an appropriate imputation method.

arrayImputehas been developed to provide an exploratory analysis of missing data and the imputation of missing values by various imputation methods. Further, it provides a comparative analysis of the imputed values to let the user choose an appropriate imputation method for the data. The advantage of arrayImpute is its user-friendly graphical interface. Therefore, the users can easily perform an analysis for the missing data by a simple click of the mouse.

Exploratory Analysis of Missing Data

arrayImpute provides various techniques to explore the missing data. It calculates missing rates for each chip and provides a bar chart to compare the chip-wise missing rates (Fig. 1a). Further, it calculates the missing rates for each gene and provides the distribution of the gene-wise missing rates (Fig. 1b).

arrayImpute also provides a heat map to display missing patterns. First, arrayImpute provides a global missing pattern plot, where the x-axis represents chips, and the y-axis represents genes (Fig. 1c). The red spots represent the missing values. From this map, we can easily find the missing patterns. In chip X04T, the missing values tend to cluster and produce several blocks. arrayImpute also produces an individual heat map for each chip, where the x and y-axes represent the two-dimensional location of the gene in the chip. Fig. 1d and 1e show the missing patterns for the chips X01T and X04T, respectively. In Fig. 1d, there are no specific patterns of red spots, which suggests that the missing pattern of X01T is random. On the other hand, in Fig. 1e, most red spots are located in the right lower corner. In this case, experimenters need to check the right lower corner of this chip for any possible artifacts of the chip.

Comparison of Imputation Methods

For the imputation of missing values, arrayImpute provides

*Corresponding author: E-mail tspark@stats.snu.ac.kr
Tel +82-2-880-8924, Fax +82-2-8830-6114
Accepted 13 August 2007

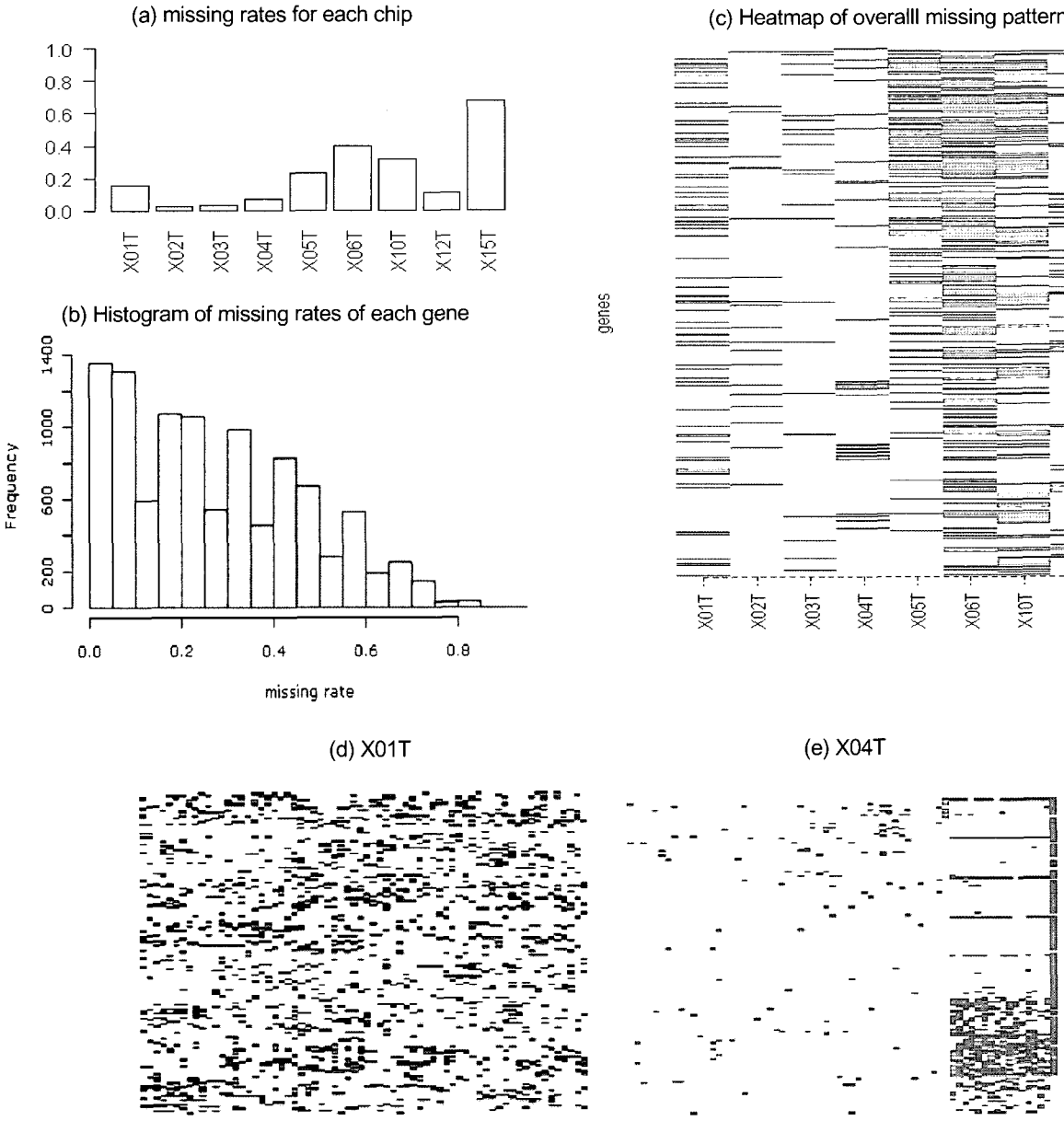

Fig. 1. Various exploratory tools in arrayImpute for estimating missing patterns. (a). Bar chart to compare the chip-wise missing rates. (b).Histogram of gene-wise missing rates. (c) Heat map to display overall missing pattern. (d)-(e) Heat map to display missing pattern for each chip.

several imputation methods such as average impute (AVE, Feten *et al*, 2005), kNN, LLS, RLSP, and BPCA. Therefore, the users can impute the missing values by these methods and simultaneously compare them to choose an appropriate imputation method. In addition, the users can upload their own imputed values from their newly developed imputation method and compare them with the imputed values provided by arrayImpute.

For an overall comparison, arrayImpute provides a scatter plot for each pair of the imputed values as well as the corresponding p-value of the pairwise t-test (Fig. 2a). arrayImpute also provides a profile plot to compare the imputed values for each gene (Fig. 2b).

In order to compare imputation methods, arrayImpute generates missing observations randomly and then imputes the missing observations using all the imputation methods available. arrayImpute computes the normalized root mean squared errors (NRMSE, Oba *et al*., 2003) for

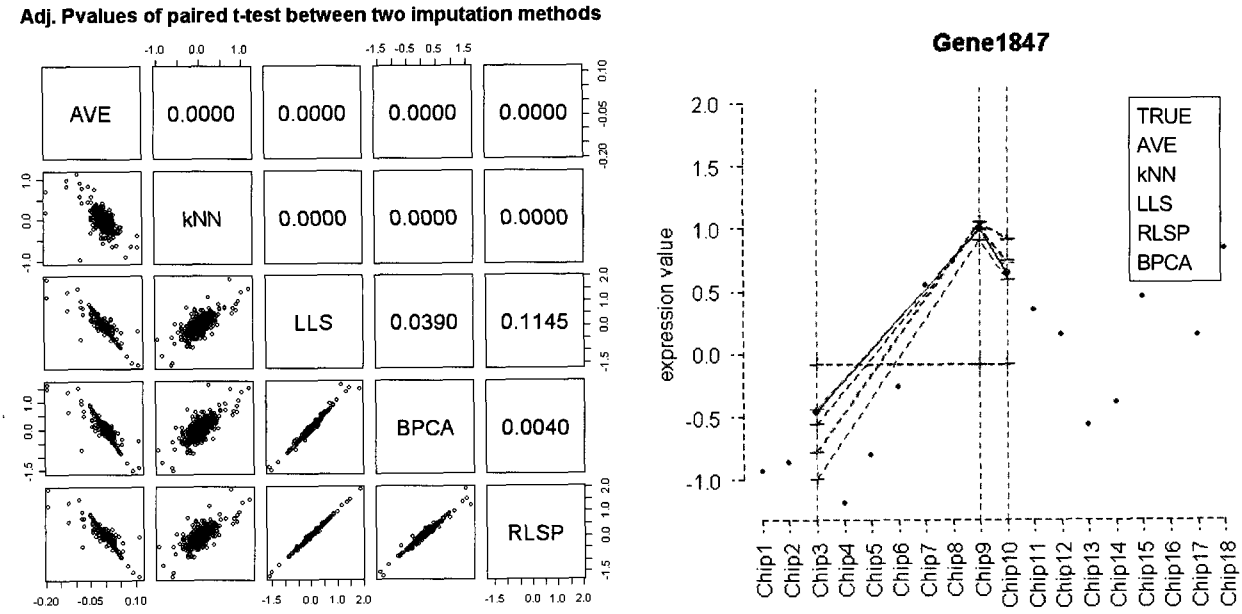

Fig. 2. Comparison of imputed values. (a) Scatter plot matrix of various imputed values with p-values from t-test. (b) Profile plot to compare imputed values for each gene. The grey dots represent non-missing data.

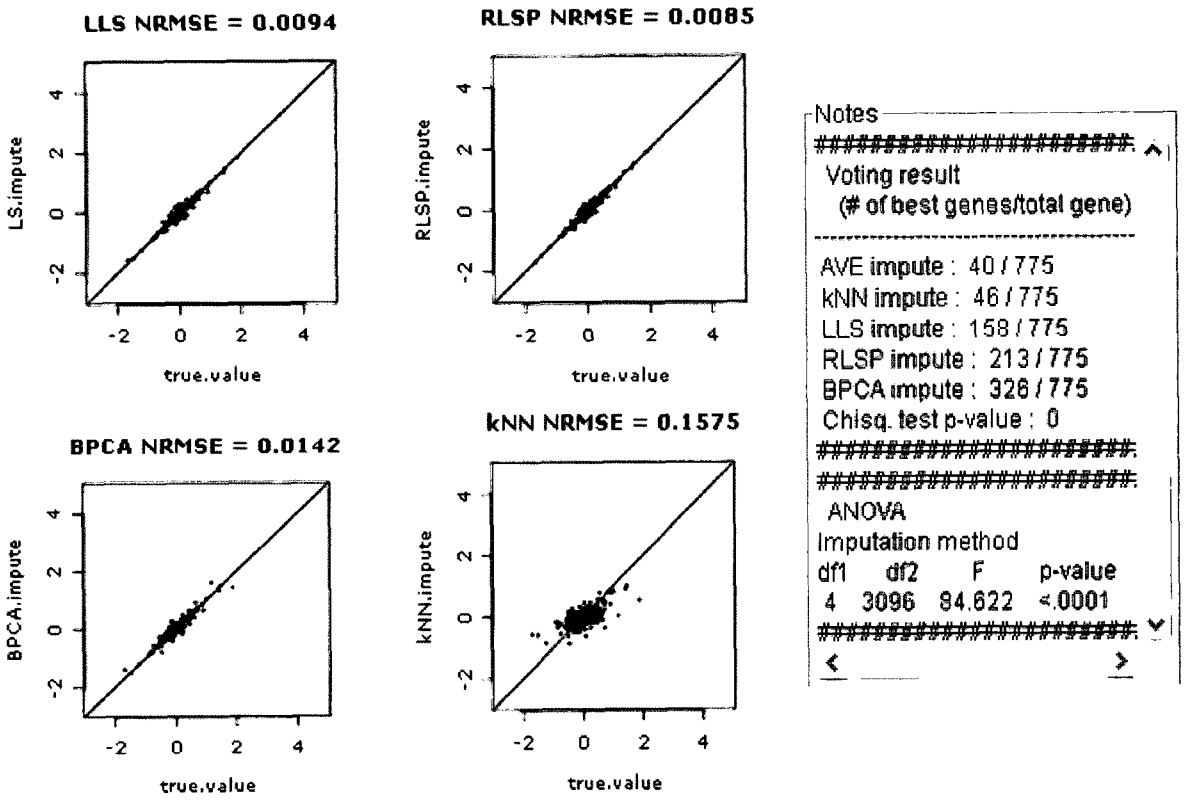

Fig. 3. Guidelines to select the best imputation method. Scatter plots with true values with NRMSE valuescompare imputed values for each gene. The grey dots represent non-missing data.

each imputation method by comparing the imputed values with the true values (Fig. 3). Even though NRMSE is the most commonly used measure for performance of imputation, NRMSE is sensitive to outliers. To overcome this limitation, arrayImpute provides another measure based on voting. For each gene, arrayImpute votes the imputation method that has the smallest absolute difference between the imputed values and the true observed values. One gene is allowed to vote for one imputation method.

arrayImpute also provides a test of significant differences between imputation methods based on the mixed effect model (Fig. 3) in which imputation method is treated as a fixed effect. All imputation methods in arrayImputeare developed under the missing at random (MAR) assumption. However, the missing patterns of some chips may be missing not at random (MNAR) (Fig.1c), and they should be treated in different ways (Scheel *et al.*, 2005). arrayImpute will be a useful tool to check whether the missing pattern is MAR or MNAR.

Implementation

This software runs on R with a couple of R packages-RGtk2 and cairoDevice-for graphical user interface.

Acknowledgments

The work was supported by the National Research Laboratory Program of Korea Science and Engineering Foundation (M10500000126).

References

Feten, G., Almoy, T., and Aastveit, A.H. (2005). Prediction of Missing Values in Microarray and Use of Mixed Models to Evaluate the Predictors *Stat Appl Genet Mol Biol*. 4, Article10.

Kim, H., Golub, G.H., and Park, H. (2005). Missing Value Estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics* 21, 187-198.

Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K, and Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* 19, 2088-2096.

Scheel, I., Aldrin, M., Glad, I., Sorum, R, Lyun, H., and Frigessi, A. . (2005). The influence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics* 21, 4272-4279.

Troyanskaya, O., Cantor, M., Sherlock, G. Brown, P., Hastie, T.,Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* 17, 520-525.

Yoon, D., Lee, E.K., and Park, T. (2007). Robust imputation method for missing values in microarray data. *BMC Bioinformatics* 8, S6.

Wang, D., Lv, Y., Guo, Z., Li, X., Li, Y., Zhu, J., Yang, D., Xu, J., Wang C., Rao, S. and Yang, B. (2006). Effects of replacing the unreliable cDNA microarray measurements on the disease classification based on gene expression profiles and functional modules. *Bioinformatics* 22, 2883-2889.