**Title:**

Reuse of imputed data in microarray analysis increases imputation efficiency

**Introduction:**

SKNN is an imputation tool, design for the imputation of gene missing entries, the tool is available as an R package.

**Contribution:**

In this paper, two novel algorithms for the gene microarray data imputation are proposed. The proposed algorithms are the extension of k-nearest neighborhood (KNN) algorithm named as sequential KNN (SKNN) and expectation maximization SKNN (EM-SKNN). The EM-SKNN is the recursive application of SKNN algorithm, both methods are compared with KNN, maximum likelihood estimation (MLE) and multiple imputations (MI) algorithms for the imputation accuracy and computation speed on different datasets.

**Working:**

In SKNN method,

1-The genes are ordered by their missing rates and the imputation was performed sequentially from a gene that had a least missing rate.

2-The sequentially imputed genes were then used in the imputation of other missing genes.

The SKNN utilizes two sets for imputation 1) complete set and 2) incomplete set. Initially, all complete and incomplete genes were placed in set 1 and 2 respectively, once a gene is fully imputed it is moved from incomplete set to complete set.

The process can be executed in parallel for all missing values of a gene. This reduces the execution time compared to the other KNN based methods

**Results:**

In this paper, five methods for the imputation of gene microarray data are discussed namely KNN, MLE, MI, SKNN, and EM-SKNN. All mentioned algorithm are compared for imputation accuracy and execution time on a different dataset with missing rates varies from (10% to 60%). The datasets consist of three types of data including Time-series, non-time series and mixed data. For the evaluation of imputation accuracy two performance measures i.e. root mean squared error (RMS) and Pearson correlation coefficients were used.

**Conclusion:**

For a gene expression with the high missing rates, the EM-SKNN is shown to outperform the SKNN, MLE, MI and KNN methods. In terms of accuracy and computational complexity, the SKNN showed the best performance compared to all other methods and

MLE as the worst.