Genomics Proteomics Bioinformatics 15 (2017) 396–404

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

APPLICATION NOTE

The Ability of Different Imputation Methods to Preserve the Significant Genes and Pathways in Cancer

Rosa Aghdam [1,*,a], Taban Baghfalaki [2,b], Pegah Khosravi [1,3,c], Elnaz Saberi Ansari [1,4,*,d]

[1] School of Biological Science, Institute for Research in Fundamental Sciences (IPM), Tehran 19395-5746, Iran
[2] Department of Statistics, Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran 14115-111, Iran
[3] Department of Physiology and Biophysics, Institute for Computational Biomedicine and Institute for Precision Medicine, Weill Cornell Medical College, New York, NY 10021, USA
[4] Institut Cochin, Inserm U1016, CNRS UMR 8104, Universit Paris Descartes UMR-S1016, F-75014 Paris, France

Received 8 January 2017; revised 18 July 2017; accepted 8 August 2017
Available online 13 December 2017

Handled by Shaoqi Rao

KEYWORDS

Gene expression;
Missing data;
Imputation method;
Significant genes;
Pathway enrichment

Abstract Deciphering important genes and pathways from incomplete **gene expression** data could facilitate a better understanding of cancer. Different **imputation methods** can be applied to estimate the missing values. In our study, we evaluated various imputation methods for their performance in preserving **significant genes** and pathways. In the first step, 5% genes are considered in random for two types of ignorable and non-ignorable missingness mechanisms with various missing rates. Next, 10 well-known imputation methods were applied to the complete datasets. The significance analysis of microarrays (SAM) method was applied to detect the significant genes in rectal and lung cancers to showcase the utility of imputation approaches in preserving significant genes. To determine the impact of different imputation methods on the identification of important genes, the chi-squared test was used to compare the proportions of overlaps between significant genes detected from original data and those detected from the imputed datasets. Additionally, the significant genes are tested for their enrichment in important pathways, using the ConsensusPathDB. Our results showed that almost all the significant genes and pathways of the original dataset can be detected in all imputed datasets, indicating that there is no significant difference in the performance of various imputation

* Corresponding authors.
E-mail: rosaaghdam@ipm.ir (Aghdam R), elnaz.saberiansari@ipm.ir (Saberi Ansari E).
[a] ORCID: 0000-0001-9045-9592.
[b] ORCID: 0000-0002-2100-4532.
[c] ORCID: 0000-0001-5071-8959.
[d] ORCID: 0000-0003-4347-7186.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2017.08.003

methods tested. The source code and selected datasets are available on http://profiles.bs.ipm.ir/softwares/imputation_methods/.

Introduction

Cancer has manifested as one of the major health problems in many countries worldwide. It is also expected to be the main cause of death in the next few years [1]. Cancer has been characterized as a heterogeneous disease, comprising various subtypes. Early diagnosis of the cancer type and stage has become essential to assist with the subsequent treatment of cancer patients [2]. With the technical advances in sequencing, it is now possible to measure the expression of all genes in a sample and stratify cancer patients into high-risk and low-risk cohorts by analyzing gene expression data using bioinformatics approaches [3].

Recognizing the genes involved in cancer is an intimidating challenge due to its importance in the molecular characterization of widely defined biological classes, which has a potential role in cancer diagnosis and treatment. The growing application of bioinformatics approaches in cancer encourages researchers to develop newer techniques involving the whole genome-based microarray. The gene expression datasets, as well as many other real-world datasets, often contain missing values, thereby affecting the inference of significant genes and the associated pathways or networks. There are many reasons for the occurrence of missing values in microarray gene expression data, *e.g.*, hybridization failures, low resolution, artifacts on the microarray, image noise, corruption, and spotting problems [4–7].

Mechanically, missing values can be classified as missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) [8]. MCAR and MAR are considered ignorable, whereas NMAR is considered non-ignorable or informative missingness. Identifying the appropriate missing mechanism and missingness rate is important for imputation algorithms [9].

For microarray gene expression datasets, there are global, local, and hybrid imputation approaches, categorized according to the information used in each case [5]. The global missing imputation methods exploit the global information of the whole dataset, whereas the local missing imputation methods use the local similarity structure of a dataset. Hybrid methods combine the two to impute missing values.

Previous studies have shown that a missingness of $\leq 1\%$ in expression data is negligible and a missingness of 1%–5% is manageable. To achieve good results in imputation for an incomplete dataset with 5%–15% missingness, it is important to use appropriate approaches. When datasets have $>15\%$ missing data, choosing imputation methods may strongly influence the results [5].

Therefore, we set out to investigate the impact of missingness factors on the imputation algorithms and evaluated the performance of 10 popular imputation methods by applying five well-known methods to acquire the significant genes from the original and imputed datasets for lung and rectal cancers. Our results indicate that similar important genes are detected in all imputed datasets, suggesting no significant difference in the performance of the imputation methods tested in terms of preserving the essential genes and pathways.

Methods

Data sources

Whole genome-based microarray data were downloaded from the Gene Expression Omnibus (GEO) database [10] with accession number GSE10072 [11] and GSE15781 [12] for lung and rectal cancer, respectively. The lung cancer dataset contains 107 samples from 58 patients with lung cancer and 49 healthy individuals, whereas the rectal cancer dataset contains 42 samples from 22 patients with rectal cancer and 20 healthy individuals. The linear model for microarray analysis (Limma) package in R [13] was used for preprocessing and analysis of the microarray data. Quantile normalization [14] is then performed to achieve the same sample distribution at each state.

Data processing for generation of missing values

The gene expression datasets often contain a small proportion of genes with missing values [5]. To generate missing values in a dataset, 5% of all genes from the original datasets were selected randomly in the first step of our study. Then, ignorable and non-ignorable types of missingness were considered at a missingness rate of 10%, 20%, and 30%, respectively. To generate ignorable missing values, the samples were randomly selected based on the three rates of missingness, and then were removed. Furthermore, to generate non-ignorable missing values, the upper or lower tails (10%, 20%, and 30%) of the data were selected, and their values were removed to ensure that the missingness depends on the actual gene expression.

Imputation methods

Ten imputation methods are considered in this study. Among them, the singular value decomposition (SVD), the Bayesian principal component analysis (BPCA), fast imputation (Fast-Imp), column-mean, column-median, gene-mean, and gene-median are global methods, whereas local least squares (LLS) and K-nearest neighbor (KNN) are local methods. Multiple imputation by chained equations and classification and regression trees (MICE-CART) is a hybrid method.

The SVD imputes missing values using the singular value decomposition and regression models [15]. The k genes similar to a target gene, which contains missing values, are detected by KNN method using a similarity metric calculated with the non-missing data. Then, the weighted average of these neighbors is calculated to impute the missing values in target gene [15]. The MICE-CART imputation method encloses MICE and CART approaches [16]. Principle component regression, an expectation–maximization (EM) algorithm, and the Bayesian estimation approach are applied in the BPCA imputation method [17]. In order to impute the missing values, a multiple regression model is applied in LLS method [18]. The EM algorithm under the multivariate normal distributional assumption is used in a Fast-Imp method to complete datasets [19]. Other