Computers in Biology and Medicine 77 (2016) 76–89

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

A global learning with local preservation method for microarray data imputation

Ye Chen [a], Aiguo Wang [a,b,*], Huitong Ding [a], Xia Que [a], Yabo Li [c], Ning An [a], Lili Jiang [d]

[a] School of Computer and Information, Hefei University of Technology, Hefei 230009, China
[b] School of Software, Hefei University of Technology, Hefei 230009, China
[c] College of Life Sciences, Lanzhou University, Lanzhou 730000, China
[d] Department of Computing Science, Umeå University, Umeå 90187, Sweden

ARTICLE INFO

Article history:
Received 13 April 2016
Received in revised form
4 August 2016
Accepted 4 August 2016

Keywords:
Missing value imputation
Microarray data
Global learning
Local preservation
Regression model

ABSTRACT

Microarray data suffer from missing values for various reasons, including insufficient resolution, image noise, and experimental errors. Because missing values can hinder downstream analysis steps that require complete data as input, it is crucial to be able to estimate the missing values. In this study, we propose a Global Learning with Local Preservation method (GL2P) for imputation of missing values in microarray data. GL2P consists of two components: a local similarity measurement module and a global weighted imputation module. The former uses a local structure preservation scheme to exploit as much information as possible from the observable data, and the latter is responsible for estimating the missing values of a target gene by considering all of its neighbors rather than a subset of them. Furthermore, GL2P imputes the missing values in ascending order according to the rate of missing data for each target gene to fully utilize previously estimated values. To validate the proposed method, we conducted extensive experiments on six benchmarked microarray datasets. We compared GL2P with eight state-of-the-art imputation methods in terms of four performance metrics. The experimental results indicate that GL2P outperforms its competitors in terms of imputation accuracy and better preserves the structure of differentially expressed genes. In addition, GL2P is less sensitive to the number of neighbors than other local learning-based imputation methods.

1. Introduction

Microarray technology can be used to simultaneously measure the expression profiles of thousands of genes under different experimental conditions [1], and microarray data analysis has been used to study disease genes [2,3], drug targets [4], and cancer subtypes [5,6]. Various machine learning and statistical analysis methods have been applied to microarray datasets for disease prediction and cancer treatment [7]. Due to the high dimensionality of gene expression profiles (which may include thousands of genes) and the small sample sizes of microarray experiments (which may be limited to tens of samples), feature extraction and feature selection are vital tools for microarray data analysis [8,9]. However, the existence of missing values in microarray datasets poses a significant problem. Previous studies have shown that most publicly available microarray datasets have rates of missing values that can reach 50% or even 95% [10]. The missing values have adverse effects on gene expression clustering and classification. Most of the existing feature selection, classification and clustering techniques require a complete dataset as input, whereas the intuitive solution of removing samples or genes with missing values results in a dramatic loss of information, especially when the removed genes play a dominant role in the biological processes of interest. Therefore, there is a practical need to precisely estimate missing values.

There are numerous human and non-human factors that can lead to missing values in microarray data, ranging from the irregular use of microarray technology and the contamination of microarray surfaces to non-specific hybridization and systematic errors in the experimental procedure [11]. For example, inappropriate manual operations can blur the fluorescence image and make it difficult to obtain accurate expression profiles. To address such issues, additional replicates of the microarray experiment can be performed. However, the high experimental costs and lack of an effective repetition scheme make this method less than ideal in practice [12]. To obtain a complete training set, some studies have proposed replacing missing values with zeros,

* Corresponding author at: School of Computer and Information, Hefei University of Technology, Hefei 230009, China.
E-mail addresses: ye1991214@126.com (Y. Chen), wangaiguo2546@163.com (A. Wang), ding_huitong@163.com (H. Ding), quexia@hfut.edu.cn (X. Que), liya19890211@126.com (Y. Li), ning.g.an@acm.org (N. An), lili.jiang@cs.umu.se (L. Jiang).

http://dx.doi.org/10.1016/j.compbiomed.2016.08.005
0010-4825/

averaging the observed values for the same gene (column mean), or averaging the observed values for the same sample (row mean) [13]. Although these existing methods are efficient and easily implemented, they fail to explore the latent data structure information such as gene co-expression and pathway relationships between genes. Thus, simplistic missing value estimation leads to large deviations from the true values.

In recent years, researchers have proposed a number of effective imputation methods, which can be broadly categorized into four groups: biological knowledge-, global learning-, local learning-, and hybrid-based methods. Biological knowledge-based methods make use of biologically validated domain knowledge, such as sample categories, gene function networks, gene regulatory networks, and gene ontology, as prior information for the target gene [14]. A major limitation of these methods is that they rely heavily on domain-specific knowledge and fail to handle situations where there is less biological knowledge available for new, under-explored cases. Global learning-based methods assume that a covariance structure exists in the dataset and use that information to estimate the missing values. Bayesian principal component analysis imputation (BPCAimpute) and singular vector decomposition imputation (SVDimpute) are examples of global learning-based methods [15]. These methods are more suitable for large microarray datasets and are sensitive to noise in the data. Particularly, they often exhibit unsatisfactory performance if similar local structures exist in the data [16]. In contrast to global learning-based methods, local learning-based methods first attempt to identify the similar local structures and then impute the missing values using genes that are similar to the target genes [17]. For example, *k*-nearest-neighbor imputation (KNNimpute) was among the earliest methods used to estimate the missing values of a target gene by weighting the values of its *k* nearest neighbors [18]. Other researchers proposed slightly different methods called least squares imputation (LSimpute) and local least squares imputation (LLSimpute), both of which introduce a regression model to build the relationships between the target gene and its neighbors and then impute the missing values using neighbors and associated regression coefficients [19,20]. Essentially, they select similar genes and build a regression equation model, but these steps are not jointly optimized, so they may not make full use of the local structure. Additionally, they are unable to use the global information provided by the data. Hybrid methods aim to take advantage of both global learning and local learning based methods. Commonly used schemes include, but are not limited to, linearly or non-linearly combining multiple imputation methods [21], integrating different imputation methods under an ensemble or semi-supervised learning framework [22], and building a pipeline using the output of an imputation method to initialize the parameter values of another imputation method [23].

Even with these state-of-the-art methods, several issues remain to be addressed. First, in local learning-based methods, the number of neighbors has to be specified. Because these methods tend to select a small fraction of the available genes as the neighbors, some potentially relevant genes may be missed. In particular, an inappropriately specified number of neighbors can dramatically degrade the performance of imputation and downstream analysis. Second, most of the existing methods process these similar genes equally and do not weight the genes based on their distances from the target gene. To address these issues, this study makes the following contributions: (1) a Global Learning with Local Preservation method (GL2P) is proposed to estimate the missing values in gene expression profiles. Specifically, GL2P imputes the missing entries by recognizing the gene neighborhood and calculating the weights of different genes. (2) A local similarity metric is proposed to fully utilize the information provided by observable data, and it is used to measure the relevance between the target gene and its similar genes. In contrast to most existing local learning-based methods that require neighbors with no missing values, GL2P relaxes this requirement and allows genes with missing values to be chosen as neighbors. (3) We propose a novel method that can automatically determine the number of similar genes to use. (4) A weighted multivariate linear regression model is trained between the target gene and the selected similar genes to estimate the missing entries of the target gene. In particular, we take gene importance into account and give more weight to genes that are more relevant to the target gene and less weight to genes that are less relevant. In addition, we exploit all genes that are similar to the target gene rather than a small fraction of the similar genes when constructing the regression model. This strategy largely mitigates the problem of loss of information, particularly for datasets with high missing rates.

The rest of this paper is organized as follows. Section 2 presents the basic notation and illustrates the core components of GL2P. In Section 3, we introduce four evaluation metrics. Extensive experimental results and an analysis of the results are presented in Section 4. Section 5 concludes the paper and considers several future research paths.

2. The proposed imputation method

In this section, after introducing the necessary notation and definitions, we present the overall framework of the proposed Global Learning with Local Preservation method. In the analysis of gene expression profiles, we usually represent the microarray data as a matrix, as shown in Fig. 1. In this study, we use $G \in \mathrm{R}^{m\times n}$ to represent the matrix, where m is the number of samples, and n is the number of genes. Specifically, we use $g_1, g_2, \ldots, g_n$ to represent the n genes and use $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_n$ ($\mathbf{g}_i \in \mathrm{R}^{m\times 1},\ 1 \le i \le n$) to indicate their vector forms. We use $s_1, s_2, \ldots, s_m$ to denote the m samples, and $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m$ ($\mathbf{s}_i \in \mathrm{R}^{1\times n},\ 1 \le i \le m$) are the corresponding vectors. That is, $G = (\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_n) = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m)^{\mathbf{T}}$. $\alpha_{i,j}$ indicates that there is a missing value at the i-th row and j-th column. For example, $\alpha_{1,2}$ means that the second gene has a missing entry in the first sample. The following section makes use of three definitions:

Definition 1 (*target gene*). In a microarray dataset consisting of expression profiles of thousands of genes, we call a gene with at least one missing value across all samples a *target gene*.

Definition 2 (*candidate gene*). In a microarray dataset consisting of expression profiles of thousands of genes, all genes excluding the target gene are called *candidate genes*. The collection of candidate genes is a candidate gene set associated with the target gene.

Definition 3 (*similar gene*). In a microarray dataset with thousands of genes, we define a gene with a similar expression pattern to the target gene as a *similar gene*. We define the collection of similar genes as the similar gene set for the target gene.

$$G=[\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_n]=\begin{bmatrix}\mathbf{s}_1^{\mathbf{T}}\\ \mathbf{s}_2^{\mathbf{T}}\\ \vdots\\ \mathbf{s}_m^{\mathbf{T}}\end{bmatrix}=\begin{bmatrix} g_{1,1} & \alpha_{1,2} & \alpha_{1,3} & \cdots & g_{1,n}\\ g_{2,1} & g_{2,2} & g_{2,3} & \cdots & g_{2,n}\\ g_{3,1} & g_{3,2} & g_{3,3} & \cdots & g_{3,n}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ g_{m,1} & \alpha_{m,2} & g_{m,3} & \cdots & g_{m,n}\end{bmatrix}$$

Fig. 1. Logical storage structure of microarray data.

For a specific target gene $\mathbf{g}_t$, its similar gene set is a subset of the candidate gene set, and both may contain missing values. To estimate missing values, the proposed imputation method GL2P works by using the following steps: 1) selecting the target gene $\mathbf{g}_t$ with minimal missing rate, 2) identifying genes similar to $\mathbf{g}_t$ from the corresponding candidate gene set, 3) calculating the local similarity measurements between $\mathbf{g}_t$ and each of its similar genes, and 4) building a regression model and using it to estimate the missing values of $\mathbf{g}_t$. After these four steps, we impute all missing entries in $\mathbf{g}_t$. Then, we choose another gene with minimal missing rate as the next target gene $\mathbf{g}_t$ and repeat the above operations until no genes remain with missing values. Algorithm 1 presents the pseudo-code of GL2P. In the next subsections, we explain the four procedures in detail.

Algorithm 1. Pseudo-code description of GL2P.

```
Input:  Dataset G with missing values
Output: Complete dataset G_c without missing values
1       while has_missing_value(G) do
2         step 1:
3       select target gene g_t with minimal missing rate;
4         step 2:
5       identify genes similar to g_t using (2) and (3);
6         step 3: //local similarity measurement
7       calculate the distance between g_t and each of its similar genes using (5);
8         step 4: //impute missing values of g_t
9       (4.1) build a regression model using (6);
10      (4.2) estimate the missing values of g_t using (10);
11      endwhile
12      return G_c;
```

2.1. Choosing the target gene

In the process of missing value imputation, there are several imputation strategies that can be used. For example, we can simply impute the missing values sequentially, from the first gene to the last. We can also estimate the missing values randomly by choosing a random gene with missing values as the target gene. Considering the fact that the previously estimated values can be used to impute the missing values of other target genes, we propose to impute the missing values in ascending order in terms of the missing rates associated with target genes. The missing rate r_i of gene $\mathbf{g}_i$ is calculated using Eq. (1),

$$r_i = \frac{l_i}{m}, \tag{1}$$

where l_i represents the number of missing values in gene $\mathbf{g}_i$, and m is the total number of samples.

Furthermore, we divide the genes in G into two groups: the incomplete gene set and the complete gene set. Assuming that n_1 genes have missing entries, the incomplete gene set $G_1 \in \mathrm{R}^{m\times n_1}$ has n_1 incomplete genes, and the complete gene set $G_2 \in \mathrm{R}^{m\times n_2}$ contains n_2 genes without missing values ($n=n_1+n_2$). For imputation, we first choose the target gene $\mathbf{g}_t$ with minimal missing rate from the incomplete gene set and estimate its missing values. After estimation, $\mathbf{g}_t$ is added to the complete gene set and excluded from the incomplete gene set.

2.2. Identifying genes similar to the target gene

The aim of this step is to determine which genes have similar expression patterns to those of the target gene from the candidate gene set. Not all candidate genes for a target gene are suitable for missing value estimation. For example, given a target gene $\mathbf{g}_1$ with a missing value in the first sample, if candidate gene $\mathbf{g}_2$ has a missing entry in the first sample, then $\mathbf{g}_2$ is not a qualified neighbor for $\mathbf{g}_1$. To filter out these unqualified and low-quality candidate genes, GL2P uses the following two constraint conditions to determine whether a candidate gene $\mathbf{g}_v$ is a similar gene to the target gene $\mathbf{g}_t$. Assuming that the indices of missing values for gene $\mathbf{g}$ are $idx(\mathbf{g})$, GL2P requires that the intersection between $idx(\mathbf{g}_t)$ and $idx(\mathbf{g}_v)$ is not empty. For example, if $\mathbf{g}_v$ has missing values in the same sample as $\mathbf{g}_t$, $\mathbf{g}_v$ provides no information to impute the missing values of $\mathbf{g}_t$. Thus, GL2P first filters out the candidate genes satisfying Eq. (2),

$$idx(\mathbf{g}_t) \cap idx(\mathbf{g}_v) = \phi \tag{2}$$

In missing value imputation, the missing rate is an important factor and is often used as an indicator of the quality of a candidate gene. GL2P takes the missing rate associated with each candidate gene into account. Specifically, it assumes that a candidate gene with more missing values contributes less to the imputation, and it filters out genes with missing rates larger than the average missing rate using Eq. (3),

$$r_v < \frac{1}{n_1}\sum_{i=1}^{n_1} r_i \tag{3}$$

In this equation, r_i is the missing rate of gene $\mathbf{g}_i$ in G_1, and n_1 is the number of genes with missing values in G_1.

2.3. Local similarity measurement

After performing the steps discussed in Subsection 2.2, we obtained a set of genes that are similar to the target gene $\mathbf{g}_t$. With this set, we can use standard distance metrics to measure the similarities between genes. Euclidean distance and Pearson correlation coefficient are two commonly used similarity metrics. However, problems arise when we directly apply these measurements to gene expression profiles. First, if two genes with similar patterns contain missing values, the existing similarity metrics cannot be used. Discarding these genes inevitably leads to a loss of information. Second, the missing rate is an important factor in measuring the confidence of calculating gene similarity, and most existing imputation methods fail to consider this. In this study, to measure the relative importance of each similar gene to a target gene, GL2P uses the following objective function (4) to calculate the distance between $\mathbf{g}_v$ and $\mathbf{g}_t$,

$$d_v = f\left(\sum_{i=1}^{|obs\cap obs0|} S\left(\mathbf{g}_t^i, \mathbf{g}_v^i\right), l_v, l_i\right) \tag{4}$$

Here, obs indicates the indices of samples without missing values for $\mathbf{g}_t$, $obs0$ indicates the indices of samples without missing values for $\mathbf{g}_v$, l_t is the number of observed values corresponding to obs in $\mathbf{g}_t$, l_v is the number of observed values corresponding to $obs\cap obs0$ in $\mathbf{g}_v$, and S is a similarity metric, such as Euclidean distance or Pearson correlation coefficient. In this study, we use Euclidean distance for similarity calculations, as expressed in Eq. (5),

$$d_v = \frac{l_v}{l_t} S(\mathbf{g}_t, \mathbf{g}_v) = \frac{l_v}{l_t}\sqrt{\sum_{i=1}^{|obs\cap obs0|}\left(\mathbf{g}_t^i - \mathbf{g}_v^i\right)^2} \tag{5}$$

As shown in the above equation, in addition to the local information for $\mathbf{g}_v$, GL2P explicitly considers its missing data rate. This helps us distinguish two genes with equal Euclidean distances but with different missing rates.

2.4. Missing value imputation with similar genes

Using the selected similar genes and their distances to the target gene $\mathbf{g}_t$, GL2P trains a weighted linear regression model considering $\mathbf{g}_t$ and all its similar genes, as shown in Eq. (6),

$$\mathbf{g}_t^{obs} = \beta_1 w_1 \mathbf{g}_1^{obs} + \beta_2 w_2 \mathbf{g}_2^{obs} + \cdots + \beta_k w_k \mathbf{g}_k^{obs} = \sum_{v=1} \beta_v \mathbf{w}_v \mathbf{g}_v^{obs} \tag{6}$$

Here, β_v represents the regression coefficients ($1 \le v \le k$), k is the number of similar genes that satisfy (2) and (3), w_v is the normalized distance between $\mathbf{g}_v$ and $\mathbf{g}_t$, which indicates the importance of $\mathbf{g}_v$ in predicting $\mathbf{g}_t$. In this study, we use a Gaussian kernel to normalize the distance:

$$w_v = exp\left(\frac{-(d_v - d_{min})}{2\sigma^2}\right) \tag{7}$$

Here, d_v is the distance between $\mathbf{g}_v$ and $\mathbf{g}_t$, d_{min} is the minimal distance between $\mathbf{g}_t$ and all of its similar genes, and σ is the kernel width, ranging from 0 to 1. Significantly, GL2P takes all similar genes into account rather than a small subset of them, which partially avoids loss of information and relieves users from having to determine the optimal number of similar genes for imputation. Moreover, a normalized distance metric can be used to normalize the calculated distances onto the same scale interval, allowing unbiased comparisons. In contrast to most of the existing methods that treat all genes equally, GL2P assigns higher weights to genes that are more similar to the target gene.

To obtain the regression coefficients of Eq. (6), the following formulas (8) and (9) are presented,

$$\min_{\beta_v}\left(\mathbf{g}_t^{obs} - \sum_{v=1}^{k} \beta_v w_v \mathbf{g}_v^{obs}\right)^2 \tag{8}$$

$$\boldsymbol{\beta} = \left[\left(\mathbf{g}_{v1}^{obs}, \mathbf{g}_{v2}^{obs}, \ldots, \mathbf{g}_{vk}^{obs}\right)\begin{bmatrix} w_1 & & \\ & w_2 & \\ & \ldots & \\ & & w_k \end{bmatrix}\right]^{+} \mathbf{g}_t^{obs} = \left(G_3^{obs} \cdot W\right)^{+} \mathbf{g}_t^{obs}, \tag{9}$$

where $(\mathbf{A})^+$ is the pseudo-inverse of matrix $\mathbf{A}$. Finally, we can estimate the missing values of $\mathbf{g}_t$ using Eq. (10),

$$\mathbf{g}_t^{miss} = \left(\beta_1, \beta_2, \ldots, \beta_k\right)\left[\mathbf{g}_{v1}^{miss}, \mathbf{g}_{v2}^{miss}, \ldots, \mathbf{g}_{vk}^{miss}\right]^{\mathbf{T}}, \tag{10}$$

where *miss* represents the indices of samples that have missing values in $\mathbf{g}_t$.

3. Evaluation metrics

To evaluate the effectiveness of GL2P in missing value imputation for microarray data, four performance metrics are used: root mean square error, Pearson correlation coefficient, conserved pairs proportion, and biomarker list concordance index. The first two metrics are statistical analysis-related metrics, and the latter two metrics evaluate the imputation methods based on biological knowledge.

3.1. Root mean square error

Root mean square error (*RMSE*) is a statistical indicator used to measure the overall deviation of estimated values from their corresponding true values. *RMSE* is defined using the following formula (11),

$$RMSE = \sqrt{\frac{1}{Z}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\hat{G}(i,j) - G_{ori}(i,j)\right]^2}, \tag{11}$$

where Z indicates the total number of missing entries in G, $\hat{G}$ represents a dataset with estimated values, and G_{ori} is a complete dataset that we use to generate G. Obviously, *RMSE* takes a value larger than 0, and a smaller *RMSE* indicates better imputation performance of the corresponding missing value estimator.

3.2. Pearson correlation coefficient

To measure how an imputation method maintains the microarray data structure, we use the Pearson correlation coefficient (12) to evaluate an imputation method.

$$correlation\ \ coefficient = \frac{cov\left(\hat{\mathbf{s}}^{\mathbf{T}}, \mathbf{s}_{ori}^{\mathbf{T}}\right)}{std\left(\hat{\mathbf{s}}^{\mathbf{T}}\right) std\left(\mathbf{s}_{ori}^{\mathbf{T}}\right)} \tag{12}$$

In this equation, $\mathbf{s}_{ori}^{\mathbf{T}}$ is a sample in G_{ori}, $\hat{\mathbf{s}}^{\mathbf{T}}$ is its corresponding sample in $\hat{G}$, $cov(\hat{\mathbf{s}}^{\mathbf{T}}, \mathbf{s}_{ori}^{\mathbf{T}})$ is the covariance between the two samples, and $std(\mathbf{s}_{ori}^{\mathbf{T}})$ represents the standard deviation of $\mathbf{s}_{ori}^{\mathbf{T}}$. The larger the Pearson correlation coefficient, the better the original data structure is maintained.

3.3. Conserved pairs proportion

Conserved Pairs Proportion (*CPP*) is a biological indicator that measures the percentage of conserved genes between the clusters of the original dataset and the imputed dataset. The number of clusters is determined by following the principle that the first ten important clusters contain at least 80% of all genes when applying hierarchical clustering with the Wards' minimum variance algorithm [24]. C_j^{ref} denotes the j-th cluster, and L_j^{ref} represents the corresponding gene list in the original dataset G_{ori}. We randomly generate a set of missing values in G_{ori} and estimate these missing values with an imputation method to obtain the imputed dataset $\hat{G}$. Here, $C_{j'}^{gen}$ and $L_{j'}^{gen}$ denote the j'-th cluster and the corresponding gene list in $\hat{G}$, respectively. *CPP* is obtained using the following formula (13),

$$CPP = \sum_{j=1}^{j=P} N_j/n, \tag{13}$$

Table 1
Description of experimental datasets.

Dataset	Original dataset (genes*samples)	Complete dataset (genes*samples)	Data type	Missing rate (%)	Genes with missing values (%)
GDS38	7680*16	5282*16	Time series	6.10	31.22
GDS1761	9706*64	8849*64	Non-time series	0.15	8.83
GDS3835	27,648*48	5070*48	Non-time series	72.25	81.63
GDS4576	22,625*9	16,052*9	Non-time series	15.63	29.05
GDS4831	24,526*22	10,523*22	Non-time series	23.75	57.09
GDS3866	10,712*28	5716*28	Mixed-time series	46.64	46.64

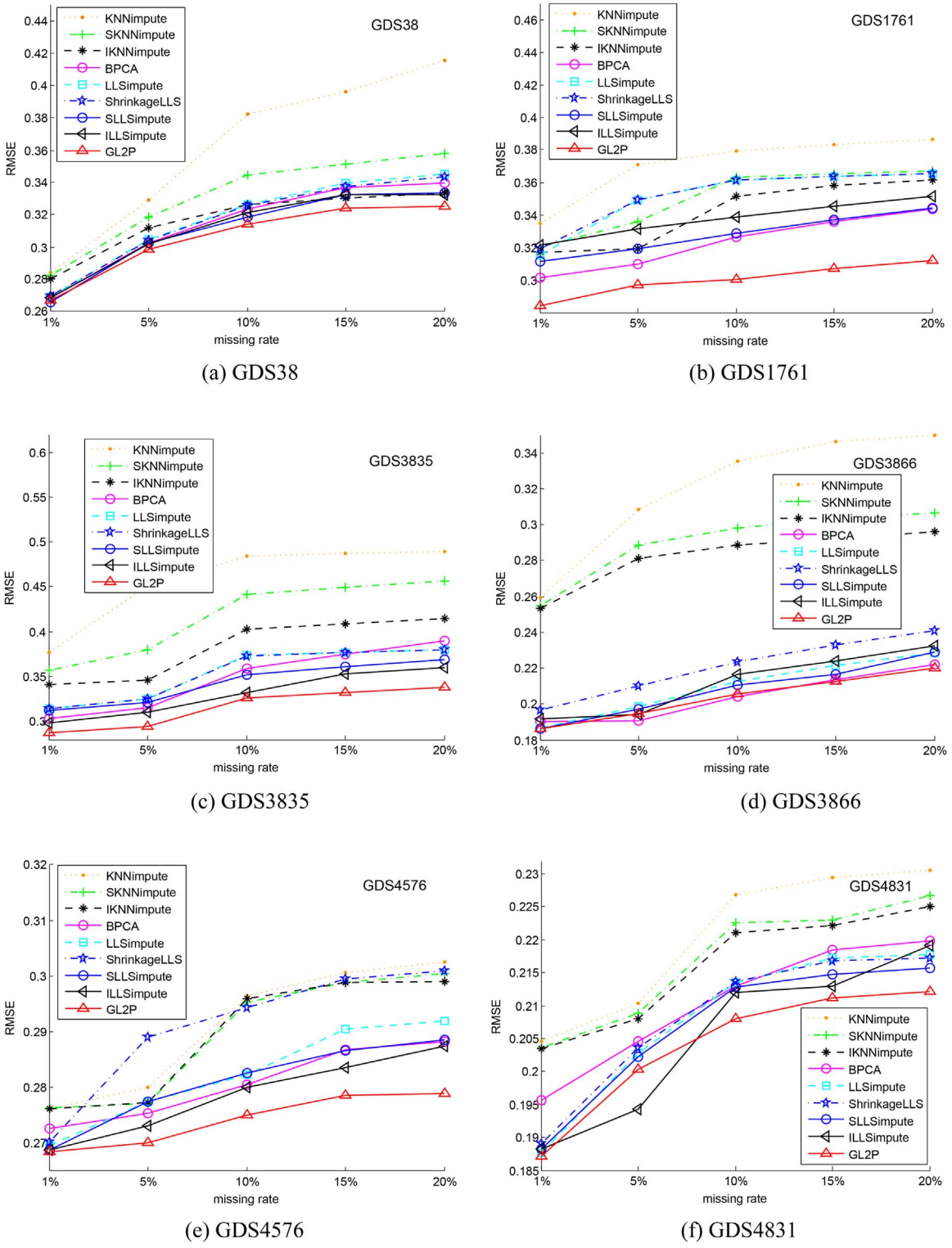
where P is the number of clusters, n is the total number of genes, and N_j is calculated using Eq. (14),

$$N_j = \max_{j'=1,\ldots,P}\left(\sum_{i\in L_j^{ref}}\sum_{i'\in L_{j'}^{gen}}\delta_{ii'}\right), \tag{14}$$

where $\delta_{ii'}$ is 1 if gene i and gene i' are equal; otherwise it is 0. *CPP* takes a maximal value of 1 if we obtain the same clusters in G_{ori} and $\hat{G}$.

3.4. Biomarker list concordance index

The biomarker list concordance index (*BLCI*) is used to measure the preservation of differentially expressed genes after imputation [25]. Suppose G_{sig} is a collection of differentially expressed genes in G_{ori}, and $\hat{G}_{sig}$ is a collection of differentially expressed genes in $\hat{G}$.



(a) GDS38 (b) GDS1761

(c) GDS3835 (d) GDS3866

(e) GDS4576 (f) GDS4831

Fig. 2. *RMSE* of different imputation methods with varying missing rates.

BLCI is calculated as follows (15):

$$BLCI\left(G_{sig}, \hat{G}_{sig}\right) = \frac{n\left(G_{sig} \cap \hat{G}_{sig}\right)}{n\left(G_{sig}\right)} + \frac{n\left(G_{sig}^{c} \cap \hat{G}_{sig}^{c}\right)}{n\left(G_{sig}^{c}\right)} - 1 \quad (15)$$

Here, $n(\cdot)$ returns the number of genes in a set, G_{sig}^{c} is the complement set of G_{sig}, and $\hat{G}_{sig}^{c}$ is the complement set of $\hat{G}_{sig}$. *BLCI* uses the SAM algorithm to detect differentially expressed genes with a false discovery rate of 5% [26]. *BLCI* has a maximal value of 1 if G_{sig} is equal to $\hat{G}_{sig}$.

4. Experimental setting and results

4.1. Experimental settings

Experiments were conducted using six publicly available microarray datasets containing both time series and non-time series data. A brief summary of the six datasets is presented in Table 1. The last column shows the ratio of incomplete genes, which indicates the varying degrees of missing values for the microarray data. All experimental datasets can be downloaded from https://www.ncbi.nlm.nih.gov/, and their brief descriptions are given below.

(1) GDS38 comes from a study of cell-cycle-regulated genes in *Saccharomyces cerevisiae*. There are 16 samples with 7680 genes, and samples were collected at different points in the cell cycle within 7 min. GDS38 is a time series dataset [27].
(2) GDS1761 is a collection of 64 samples from 60 cell lines from a variety of tissues and organs. Each sample consists of 9706 genes. It is a non-time series dataset [28].
(3) GDS3835 comes from the analysis of three pure and one hybrid species of *Drosophila* during four different developmental stages (larval stage, early pupal stage, late pupal stage, and adult stage). Each species repeats these stages three times; therefore, there are 48 samples with 27,648 genes in each sample [29].
(4) GDS4576 consists of 9 samples from a study of young adults after a 4-h exposure to live *Candida albicans* and heat-killed *Candida albicans*. It is a non-time series dataset, and each sample has 22,625 genes [30].
(5) GDS4831 comes from the study of the ubiquitin modulator depletion effect on hepatocellular carcinoma cell lines. It is a non-time series dataset with 22 samples and 24,526 genes [31].
(6) GDS3866 comes from a study of yeast response to sudden oxygen depletion. There are two subsets in the 28 samples: the first 14 samples were taken at time points of 0 h, 0.2 h, 1 h, 3 h, 8 h, 24 h, and 72 h after sudden oxygen level shifts from 1% to 0%, and the rest of the samples were taken at time points of 0 h, 0.2 h, 1 h, 3 h, 8 h, 24 h, 79 h after sudden oxygen level shifts from 20% to 0%. It is a mixed-time series dataset with 10,712 genes [32].

To evaluate the effectiveness of GL2P with different missing rates, we first obtained complete microarray data without missing values by deleting genes with missing entries from the original data, and we then randomly introduced artificial missing values to the complete datasets with missing rates of 1%, 5%, 10%, 15%, and 20%, respectively. For comparison experiments, we compared our proposed GL2P method with eight other state-of-the-art methods, including KNNimpute [18], SKNNimpute [33], IKNNimpute [34], LLSimpute [20], ShrinkageLLS [35], SLLSimpute [36], ILLSimpute [37], and BPCA [15]. The first seven methods are based on the *k*-nearest-neighbor approach and require selecting a value for the parameter *k*. The last method, BPCA, is a global learning-based method, and it uses Bayesian theory and principal component analysis for missing value imputation. Specifically, SKNNimpute and IKNNimpute are improved versions of KNNimpute that use sequential and iterative imputation strategies; similarly,

Table 2
Experimental results of significance test on GDS38.

Missing rate (%)	KNNimpute	SKNNimpute	IKNNimpute	BPCA	LLSimpute	ShrinkageLLS	SLLSimpute	ILLSimpute
1	<	<	<	=	=	=	=	=
5	<	<	<	=	<	<	=	=
10	<	<	<	<	<	<	<	<
15	<	<	<	<	<	<	<	<
20	<	<	<	<	<	<	<	<

Table 3
Experimental results of significance test on GDS1761.

Missing rate (%)	KNNimpute	SKNNimpute	IKNNimpute	BPCA	LLSimpute	ShrinkageLLS	SLLSimpute	ILLSimpute
1	<	<	<	<	<	<	<	<
5	<	<	<	<	<	<	<	<
10	<	<	<	<	<	<	<	<
15	<	<	<	<	<	<	<	<
20	<	<	<	<	<	<	<	<

Table 4
Experimental results of significance test on GDS3835.

Missing rate (%)	KNNimpute	SKNNimpute	IKNNimpute	BPCA	LLSimpute	ShrinkageLLS	SLLSimpute	ILLSimpute
1	<	<	<	<	<	<	<	<
5	<	<	<	<	<	<	<	<
10	<	<	<	<	<	<	<	=
15	<	<	<	<	<	<	<	<
20	<	<	<	<	<	<	<	<

SLLSimpute and ILLSimpute are based on LLSimpute. ShrinkageLLS incorporates a shrinkage algorithm to adjust the regression coefficients. In contrast, GL2P can automatically determine how many similar genes to use and considers the relative importances of similar genes in estimating the missing values.

In GL2P, the kernel width σ in Eq. (7) is the only parameter to be set. In this study, we adopt the following strategy to determine the value of σ. GL2P generates candidates with the kernel width value σ ranging from 0 to 1 with a step size of 0.1 (σ=0.1, 0.2, …, 1). For each assignment, we introduce an artificial missing value p in the observed values of the target gene, and according to other observable values of the target gene, we can obtain an estimated value p_0 for p using Eqs. (9) and (10). Then, for each artificial missing value in the target gene, we can obtain its corresponding estimated value. Finally, we calculate their deviations from the true observed values using Eq. (16).

$$\text{deviation}=\sum_i \left[\alpha_i - \hat{\alpha}_i(\sigma)\right]^2 \tag{16}$$

In this equation, α_i is the i-th observed value for the target gene, and $\hat{\alpha}_i(\sigma)$ is the corresponding estimated value for a given value of σ. Thus, among all the deviations for each assignment of σ, we select the optimal parameter value with minimal deviation.

4.2. Experimental results and analysis

4.2.1. Accuracy of the imputation methods

In this study, for each missing rate, we repeated the experiments ten times for each imputation method and calculated mean values across replicates. Fig. 2 presents the experimental results of nine imputation methods on six microarray datasets. Fig. 2 shows that the *RMSE* increases with the missing rates for all imputation methods. This result is reasonable, as a higher percentage of missing values indicates loss of more useful information and creates greater challenges for imputation. Clearly, GL2P consistently outperforms the other eight competing methods, particularly in cases with high missing data rates. Thus, the experimental results demonstrate the effectiveness of GL2P for microarray data imputation.

Among the *k*-nearest-neighbor-based methods, KNNimpute gives the worst imputation accuracy. SKNNimpute outperforms KNNimpute because it sequentially imputes the missing values and can make use of previously estimated values. IKNNimpute works in an iterative way to improve the quality of imputation and outperforms KNNimpute and SKNNimpute. For the least squares-based methods, including LLSimpute, ShrinkageLLS, SLLSimpute, and ILLSimpute, Fig. 2 shows that ShrinkageLLS does not consistently exhibit better performance than LLSimpute; for example, it shows worse performance than LLSimpute on GDS4576. In comparison with LLSimpute, both SLLSimpute and ILLSimpute show improvement in *RMSE*. We also observe that the least squares-based methods tend to achieve better imputation accuracy than *k*-nearest-neighbor based methods. This indicates the superiority of the least-squares principle over nearest-neighbor techniques. For the global learning-based method, BPCA shows similar performance to the least squares-based methods in most cases. Compared with GL2P, BPCA has worse imputation performance with the exception that BPCA achieves similar imputation performance to that of GL2P on GDS3866, possibly because a covariance structure exists in GDS3866. However, as the missing rate increases, the latent covariance structure may be destroyed, which limits the power of BPCA for effective imputation.

We used a *t*-test with a significance interval of 95% to determine whether there was any difference between GL2P and its competitors in terms of *RMSE*. In our experiments, differences were considered significant if the *p*-value was less than 0.05. Tables 2–7 present the experimental results for the six datasets. The symbol "<" in the tables indicates that GL2P exhibits a lower *RMSE*, ">"indicates that GL2P performs worse, and "=" indicates that there is no significant difference between GL2P and the competing method. A smaller *RMSE* reflects a lower deviation of imputation. According to the results in Tables 2–7, we observe that when the missing rate reaches 10%, GL2P gives a significantly smaller *RMSE* than the other eight imputation methods. Specifically, the *RMSE* of GL2P is significantly smaller than the values of KNNimpute, SKNNimpute and IKNNimpute on all experimental datasets. Additionally, GL2P consistently has a smaller *RMSE* than the other imputation methods on GDS1761 and GDS3835, regardless of the missing data rate. Furthermore, on GDS3866, BPCA shows results close to those of GL2P when the missing data rate is less than 15%, but GL2P outperforms it as the missing rate reaches 20%.

4.2.2. Preservation of data structure

To show the effectiveness of different imputation methods in preserving the original data structure, we conducted experiments on each dataset with a representative 5% missing rate. Fig. 3 presents the experimental results of the preservation of data structure for the nine imputation methods. In Fig. 3, we can see that GL2P consistently outperforms the other eight imputation methods on all experimental datasets, demonstrating that GL2P can better

Table 5
Experimental results of significance test on GDS3866.

Missing rate (%)	KNNimpute	SKNNimpute	IKNNimpute	BPCA	LLSimpute	ShrinkageLLS	SLLSimpute	ILLSimpute
1	<	<	<	<	=	<	=	<
5	<	<	<	>	=	<	=	=
10	<	<	<	=	<	<	<	<
15	<	<	<	=	<	<	<	<
20	<	<	<	<	<	<	<	<

Table 6
Experimental results of significance test on GDS4576.

Missing rate (%)	KNNimpute	SKNNimpute	IKNNimpute	BPCA	LLSimpute	ShrinkageLLS	SLLSimpute	ILLSimpute
1	<	<	<	<	=	=	=	=
5	<	<	<	<	<	<	<	<
10	<	<	<	<	<	<	<	<
15	<	<	<	<	<	<	<	<
20	<	<	<	<	<	<	<	<

Table 7
Experimental results of significance test on GDS4831.

Missing rate (%)	KNNimpute	SKNNimpute	IKNNimpute	BPCA	LLSimpute	ShrinkageLLS	SLLSimpute	ILLSimpute
1	<	<	<	<	=	<	=	=
5	<	<	<	<	<	<	<	>
10	<	<	<	<	<	<	<	<
15	<	<	<	<	<	<	<	<
20	<	<	<	<	<	<	<	<

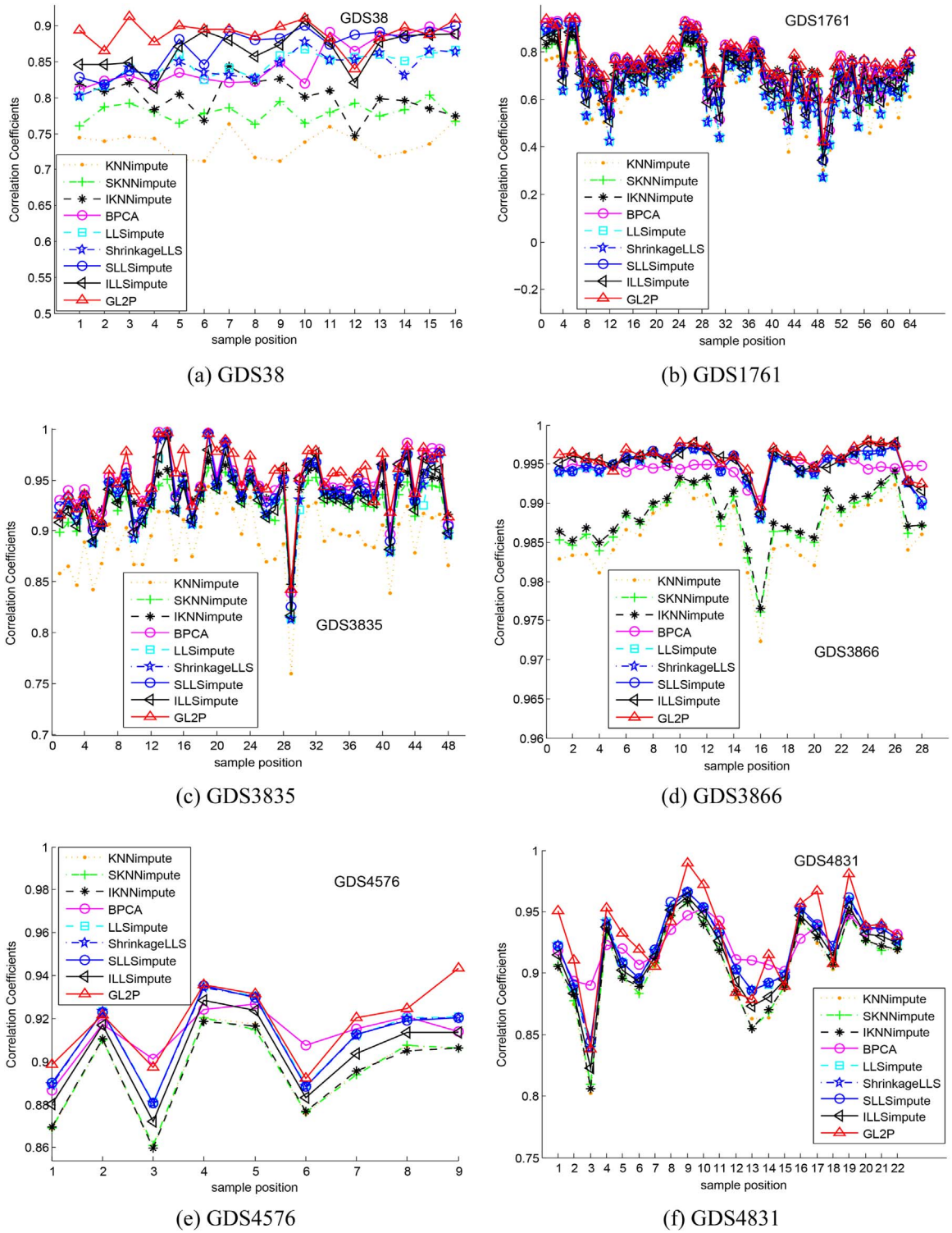


Fig. 3. Comparison of different imputation methods on correlation coefficients.

recover the missing values. For example, on GDS3866, the correlation coefficients of GL2P range from 0.97 to 1. For the nearest-neighbor-based methods, KNNimpute, SKNNimpute and IKNNimpute have worse performance than the other methods. For BPCA, the performance is comparable to that of the least squares-based methods but worse than that of GL2P.

Interestingly, considering the results of *RMSE* and the preservation of data structure, we can see that methods with higher *RMSE* may better preserve the data structure. For example, BPCA has a higher *RMSE* than ILLSimpute on GDS4576, as shown in Fig. 2b, but BPCA exhibits larger correlation coefficients than ILLSimpute, as shown in Fig. 3.e. This is because *RMSE* reflects the overall degree of deviation of imputation and fails to show the details, whereas correlation coefficients measure performance at the sample level. Consequently, this motivates us to use these two metrics together.

4.2.3. Conserved pairs proportion

In addition to *RMSE* and the preservation of data structure, experiments involving the conserved pairs proportion (*CPP*) are

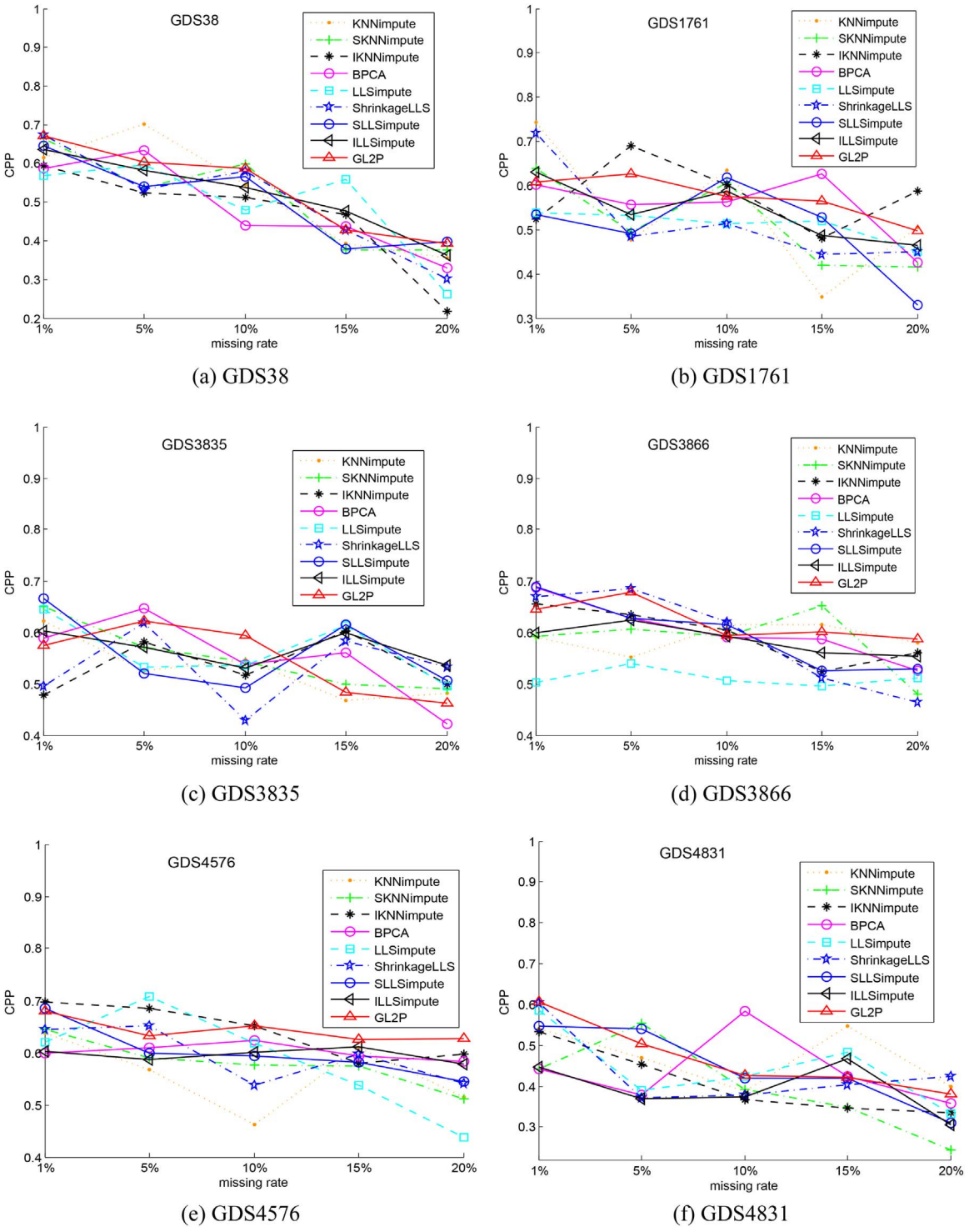


Fig. 4. Comparison of different imputation methods on conserved pairs proportion.

used to evaluate missing value imputation methods from the standpoint of biological gene clustering. A higher *CPP* indicates better imputation performance for the corresponding estimator. Fig. 4 presents the experimental results of *CPP* for different imputation methods on the experimental datasets with varying degrees of missing rates.

In Fig. 4, the *CPP* of six datasets ranges from 30% to 70%, showing that even a small missing rate can have a great impact on gene clustering, which shows that missing values in gene expression profiles can greatly affect the downstream biological analysis. A higher missing data rate has a greater effect on gene clustering due to information loss. We can also see that *CPP* slowly declines with increasing missing data rates. Unexpectedly, none of the nine methods consistently shows a better *CPP* than the others. For example, on GDS38, KNNimpute has the best *CPP* at a 5% missing rate but has the worst *CPP* at a 15% missing rate. Similar results were obtained in another study evaluating the impact of missing value methods on clustering [38]. This similarity is mainly because a smaller deviation of the estimated values from their true values has a greater impact on the results of hierarchical clustering and, thus, it leads to a smaller *CPP*. Compared with the other competing methods, however, GL2P tends to have a stable *CPP* as the missing data rate increases. This result shows that GL2P is less sensitive to perturbations of the gene expression values from their true values and can better preserve the clustering results.

4.2.4. Biomarker list concordance index

In this section, *BLCI* is used to evaluate the preservation of

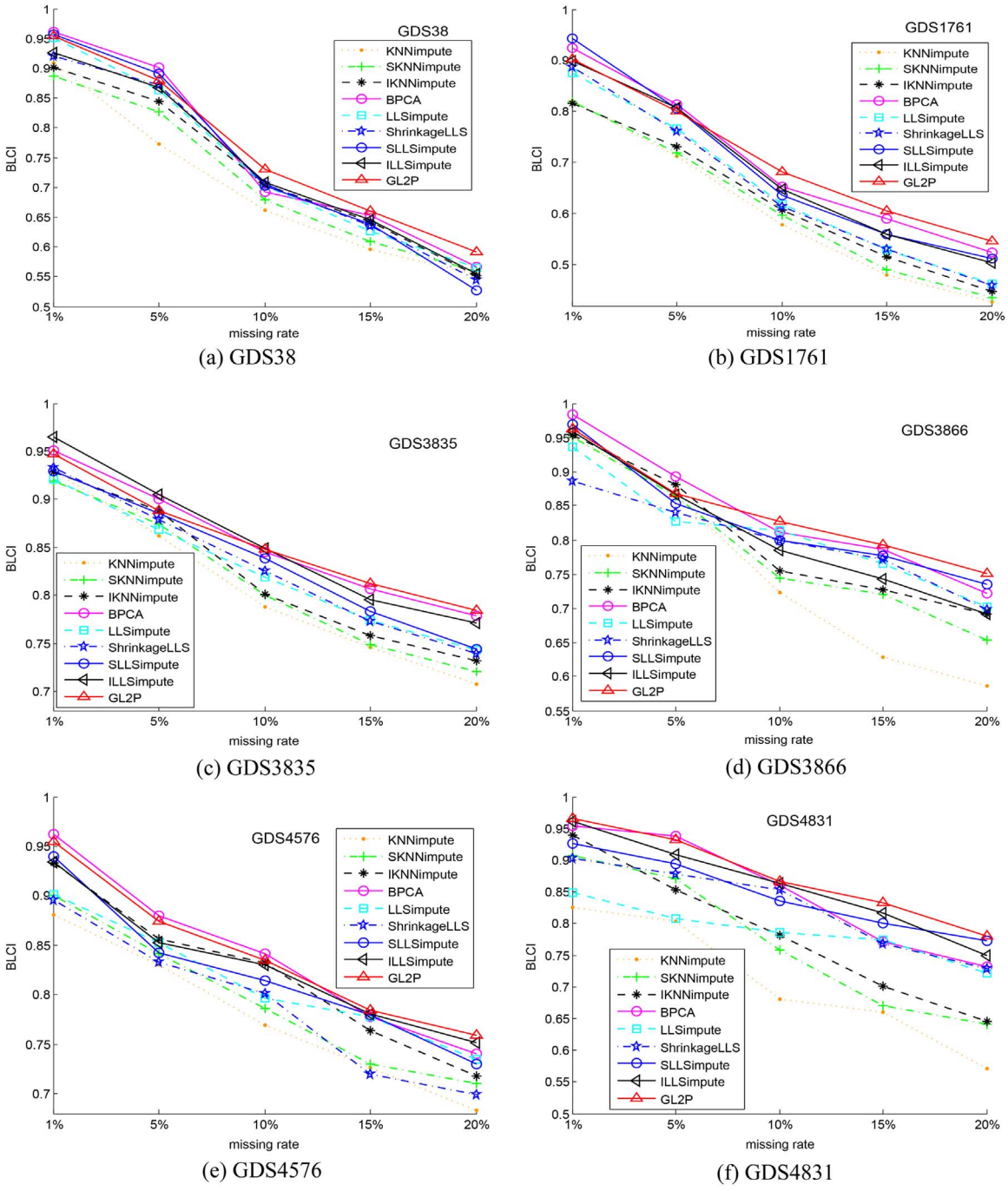


Fig. 5. Comparison of different imputation methods on biomarker list concordance index.

differentially expressed genes for different imputation methods. Fig. 5 shows a clear trend that applies to all experimental datasets for the nine imputation methods: a larger missing rate leads to a lower *BLCI*, and *BLCI* decreases quickly as the missing data rate increases. For example, at a missing data rate of 1%, most imputation methods result in a *BLCI* as high as 90%, even as high as 98%; but at a missing data rate of 20%, the *BLCI* decreases to 40%. It is reasonable that larger missing rates are accompanied by the loss of more useful information. In contrast to *CPP*, when *BLCI* is calculated, only the differentially expressed genes are considered, whereas *CPP* is calculated by considering all genes divided into different clusters. Overall, GL2P and BPCA give better *BLCI* results than other imputation methods, and BPCA yields a slightly higher *BLCI* than that of GL2P at low missing data rates. At higher missing data rates, GL2P performs better. This result demonstrates the superiority of GL2P, which attempts to use as much information as possible from the observable data. Additionally, the least-squares based methods generally have a larger *BLCI* than the nearest-

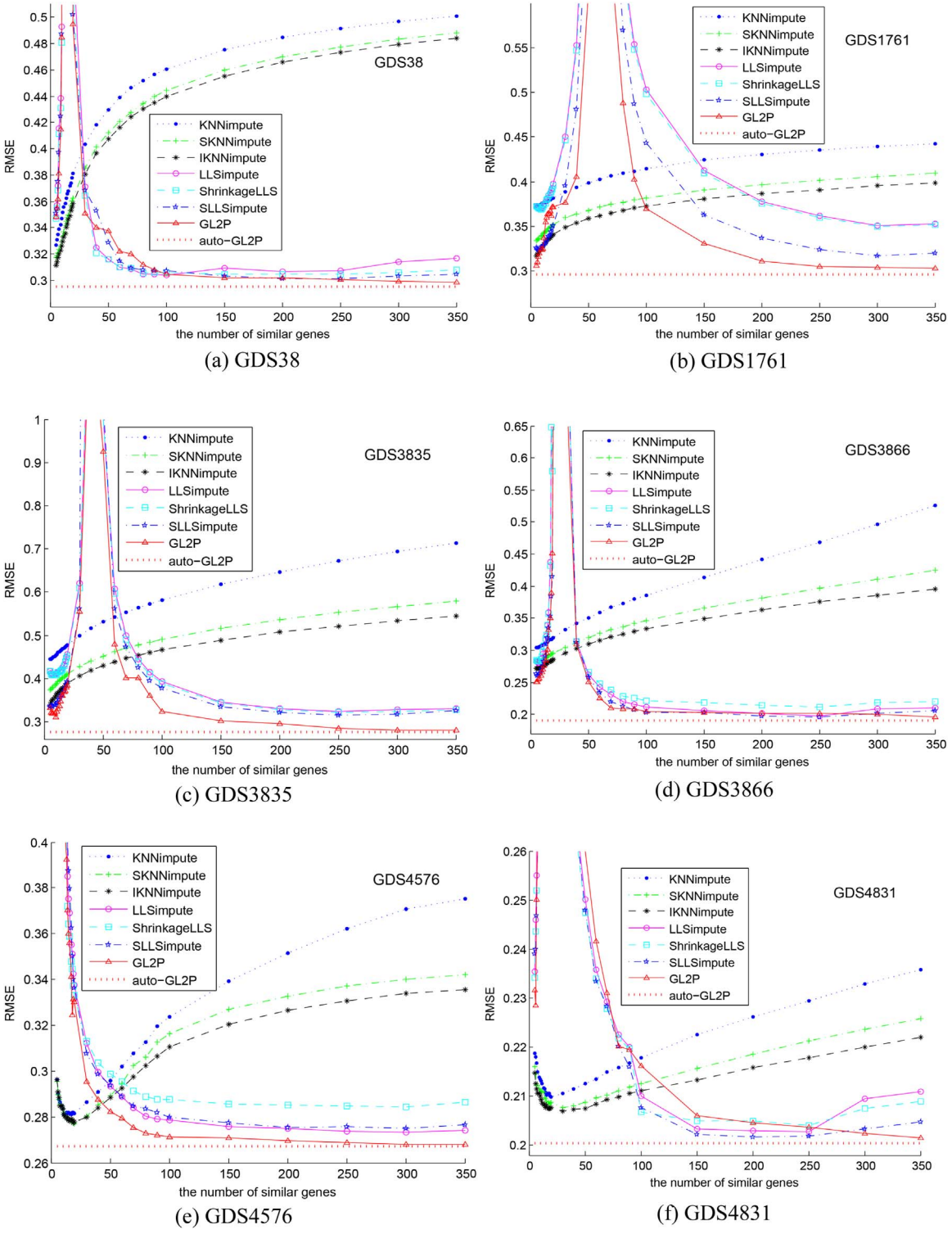


Fig. 6. Parameter sensitivity analysis to the number of similar genes.

neighbor based methods, which is consistent with the experimental results for *RMSE* and preservation of data structure.

4.2.5. Sensitivity to the number of similar genes

For local learning-based methods, including KNNimpute, SKNNimpute, IKNNimpute, LLSimpute, ShrinkageLLS, SLLSimpute, and GL2P, the number of neighbors used in estimating the missing values plays a significant role in determining the imputation quality. In this section, we conduct extensive experiments to investigate the impact of the number of neighbors on root mean square error at varying missing data rates. We set up the experiments with a missing rate of 5%, and the number of neighbors ranged from 1 to 350. Fig. 6 presents the results of these tests.

In Fig. 6, the *x*-axis indicates the number of neighbors used in estimating the missing values, and the *y*-axis indicates the performance of each imputation method in terms of *RMSE*. The red-dashed line "auto-GL2P" represents the results of GL2P that can automatically determine the number of neighbors to use, and "auto-GL2P" is used as a baseline method in these experiments. GL2P in Fig. 6 corresponds to the method with a pre-assigned

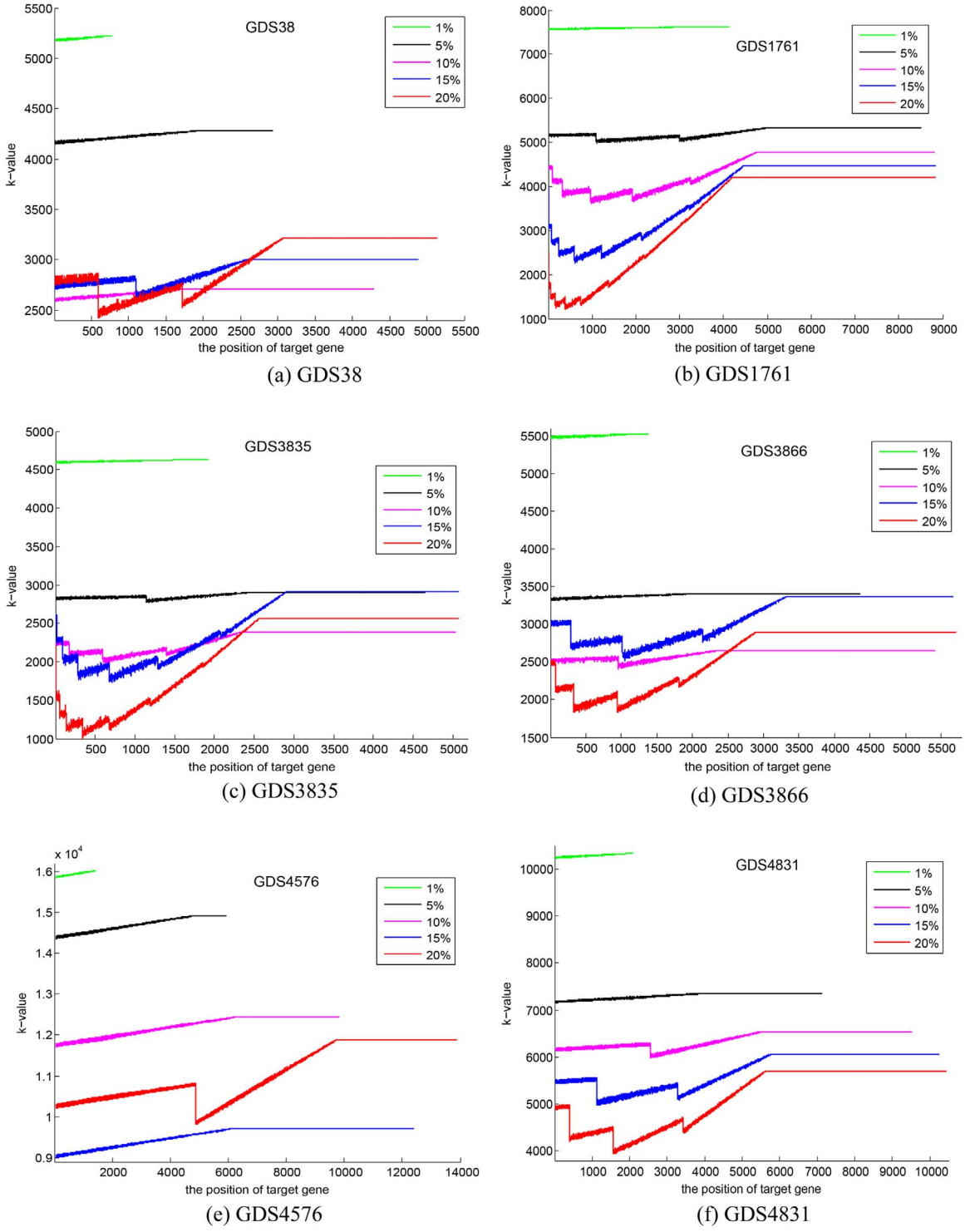


Fig. 7. The distribution of the number of similar genes for each target gene.

number of neighbors. In Fig. 6, we observe that for the least squares-based methods, *RMSE* rises quickly when k is close to the number of samples in the dataset, and *RMSE* decreases with an increase in the number of considered neighbors. One possible explanation is that when the number of neighbors considered equals the number of samples, the solution to the pseudo-inverse of matrix $\mathbf{A}$ in Eq. (9) is not fully optimized. Therefore, in practical use, a larger number of neighbors are preferred for the least squares-based methods. Regarding the nearest-neighbor-based methods, KNNimpute, SKNNimpute and IKNNimpute achieve the lowest *RMSE* when k is between 10 and 15. However, with increasing k, the nearest-neighbor based methods become worse in terms of *RMSE* because these methods use the distance metric to measure the relative importance of each similar gene rather than using a regression model to identify important genes. When more similar genes are considered, they add noise to the nearest-neighbor based methods, which inevitably leads to poor performance.

As shown in Fig. 6, GL2P can yield a smaller *RMSE* as k increases for each of the six datasets. The reason why GL2P shows better performance than the least squares methods is that GL2P considers not only the complete genes but also the incomplete genes as potential similar genes for the target gene, which enables GL2P to preserve as much useful information as possible and allows it to exclude noisy genes. Making use of the incomplete genes in the process of identifying similar genes helps improve the performance of GL2P. Furthermore, auto-GL2P consistently outperforms competing methods on all experimental datasets, and it can automatically determine the number of similar genes to use.

As mentioned above, GL2P can automatically determine the number of similar genes to use for imputing missing values for a target gene. We plotted the distribution of the number of similar genes for each target gene, as shown in Fig. 7. The x-axis represents the order in which the corresponding gene is imputed. For example, a value of 1 means that the gene is the first one to be imputed, and a value of 10 means that the corresponding gene is the tenth one to be imputed. The y-axis represents the number of genes associated with the target genes, and a higher value means that the target gene has more similar genes. Fig. 7 indicates that the number of similar genes for different target genes is not always the same, although other local learning-based methods use the same number of similar genes. Therefore, GL2P can adaptively determine the optimal number of neighbors for different target genes. In addition, we observe that the genes imputed in the late phase have more neighbors than those in earlier phases because GL2P uses a sequential imputation strategy that makes use of observable information.

Overall, extensive experiments demonstrate that in terms of root mean square error, GL2P outperforms eight state-of-the-art imputation methods, including three nearest-neighbor-based methods (KNNimpute, SKNNimpute, and IKNNimpute), four least squares-based methods (LLSimpute, SLLSimpute, ILLSimpute, and ShrinkageLLS), and one global learning-based method (BPCA). Moreover, GL2P better maintains the structure of the original data structure and better preserves the differentially expressed genes. In addition, the experimental results show that GL2P is less sensitive to the number of neighbors than other local learning-based imputation methods. Finally, GL2P has the ability to adaptively determine the number of neighbors to use for different target genes so that it is not necessary to manually assign a constant number of neighbors or to determine the optimal number via cross-validation techniques.

5. Conclusions

It is critical to estimate missing values in gene expression profiles to enable downstream analyses, such as gene selection, gene clustering, and cancer diagnosis. This study proposes a global learning with local preservation imputation method, named GL2P, to estimate the missing values in microarray datasets. GL2P follows a sequential scheme by selecting the target gene with the smallest missing rate each time rather than randomly selecting a gene, which allows it to use the previously estimated values when dealing with other target genes in the later stages. Moreover, we designed two specific constraint conditions for automatically determining the number of similar genes for each target gene, which relieves users from having to set the optimal number of neighbors. To evaluate the effectiveness of GL2P, we conducted extensive experiments on six publicly available microarray datasets and compared GL2P to eight state-of-the-art imputation methods, including seven local learning-based methods and one global learning-based method, in terms of four performance metrics. The experimental results show that GL2P outperforms the other methods in terms of imputation accuracy and effectively maintains the structure of differentially expressed genes. The experimental study further indicates that GL2P is less sensitive to the number of neighbors.

In the future, there are several research paths worth exploring. First, the distribution of missing values is an important factor for evaluating an imputation method, and this study has considered only a random set of missing values. Therefore, investigating the relationship between the performance of an imputation method and the distribution of missing values is an interesting topic. Second, GL2P could be further applied to other fields that suffer from missing values, including proteomics and clinical data analysis.

A conflict of interest statement

None declared.

Acknowledgments

This work was supported in part by the China Postdoctoral Science Foundation (No. 2016M592046), the International S&T Cooperation Program of China (No. 2014DFA11310), and the “111 Project” of the Ministry of Education and State Administration of Foreign Experts Affairs (No. B14025). The authors are very grateful to the anonymous reviewers for their constructive comments and suggestions for the improvement of this research.

References

[1] D.J. Lockhart, E.A. Winzeler, Genomics, gene expression and DNA arrays, Nature 405 (2000) 827–836.
[2] M.S. Inkeles, P.O. Scumpia, W.R. Swindell, D. Lopez, R.M. Teles, T.G. Graeber, R. L. Modlin, Comparison of molecular signatures from multiple skin diseases identifies mechanisms of immunopathogenesis, J. Investig. Dermatol. 135 (2015) 151–159.
[3] R.S. Fehrmann, J.M. Karjalainen, M. Krajewska, H.J. Westra, D. Maloney, A. Simeonov, et al., Gene expression analysis identifies global gene dosage sensitivity in cancer, Nat. Genet. 47 (2015) 115–125.
[4] W. Wang, N.G. Iyer, H.T. Tay, Y. Wu, T.K. Lim, L. Zheng, P.K. Chow, Microarray profiling shows distinct differences between primary tumors and commonly used preclinical models in hepatocellular carcinoma, BMC Cancer 15 (2015) 828.
[5] O.A. Stefansson, S. Moran, A. Gomez, S. Sayols, C. Arribas-Jorba, J. Sandoval, J. Eyfjord, A DNA methylation-based definition of biologically distinct breast cancer subtypes, Mol. Oncol. 9 (2015) 555–568.
[6] E. Cuyàs, B. Martin-Castillo, B. Corominas-Faja, A. Massaguer, J. Bosch-Barrera, J.A. Menendez, Anti-protozoal and anti-bacterial *anti*biotics that inhibit protein synthesis kill cancer subtypes enriched for stem cell-like properties, Cell Cycle 14 (2015) 3527–3532.

[7] J.E. Mirus, Y. Zhang, C.I. Li, A.E. Lokshin, R.L. Prentice, S.R. Hingorani, P. D. Lampe, Cross-species antibody microarray interrogation identifies a 3-protein panel of plasma biomarkers for early diagnosis of pancreas cancer, Clin. Cancer Res. 21 (2015) 1764–1771.
[8] M. Lenz, F.J. Müller, M. Zenke, A. Schuppert, Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data, Sci. Rep. 6 (2016) 25696.
[9] A. Wang, A. Ning, G. Chen, L. Lian, G. Alterovitz, Improving PLS–RFE based gene selection for microarray data classification, Comput. Biol. Med. 62 (2015) 14–24.
[10] M.C. Souto, P.A. Jaskowiak, I.G. Costa, Impact of missing data imputation methods on gene expression clustering and classification, BMC Bioinform. 16 (2015) 64.
[11] M.N. Arbeitman, E.E. Furlong, F. Imam, E. Johnson, B.H. Null, B.S. Baker, M. A. Krasnow, M.P. Scott, R.W. Davis, K.P. White, Gene expression during the life cycle of Drosophila melanogaster, Science 297 (2002) 2270–2275.
[12] A.J. Butte, J. Ye, G. Niederfellner, K. Rett, H.U. Häring, M.F. White, I.S. Kohane, Determining significant fold differences in gene expression analysis, in: Proceedings of the Pacific Symposium on Biocomputing (PSB), February 2001, pp. 6–17.
[13] R. Jörnsten, H.Y. Wang, W.J. Welsh, M. Ouyang, DNA microarray data imputation and significance analysis of differential expression, Bioinformatics 21 (2005) 4155–4161.
[14] Y. Yang, Z. Xu, D. Song, Missing value imputation for microRNA expression data by using a GO-based similarity measure, BMC Bioinform. 17 (2016) 10.
[15] S. Oba, M.A. Sato, I. Takemasa, M. Monden, K. Matsubara, S. Ishii, A. Bayesian, missing value estimation method for gene expression profile data, Bioinformatics 19 (2003) 2088–2096.
[16] A. Suyundikov, J.R. Stevens, C. Corcoran, J. Herrick, R.K. Wolff, M.L. Slattery, Accounting for dependence induced by weighted KNN imputation in paired samples, motivated by a colorectal cancer study, PLoS One 10 (2015) e0119876.
[17] G. Tutz, S. Ramzan, Improved methods for the imputation of missing data by nearest neighbor methods, Comput. Stat. Data Anal. 90 (2015) 84–99.
[18] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics 17 (2001) 520–525.
[19] T.H. Bø, B. Dysvik, I. Jonassen, LSimpute: accurate estimation of missing values in microarray data with least squares methods, Nucleic Acids Res. 32 (2004) e34.
[20] H. Kim, G.H. Golub, Missing value estimation for DNA microarray gene expression data: local least squares imputation, Bioinformatics 21 (2005) 187–198.
[21] S. Chattopadhyay, C. Das, S. Bose, A novel biclustering based missing value prediction method for microarray gene expression data, in: Proceedings of the 2015 International Conference on Man and Machine Interfacing (MAMI), IEEE, December 2015, pp. 1–6.
[22] H. Li, C. Zhao, F. Shao, G.Z. Li, X. Wang, A hybrid imputation approach for microarray missing value estimation, BMC Genom. 16 (2015) s1.
[23] F. Shi, D. Zhang, J. Chen, H.R. Karimi, Missing value estimation for microarray data by Bayesian principal component analysis and iterative local least squares, Math. Probl. Eng. 16 (2013) 301–312.
[24] A.G. Brevern, S. Hazout, A. Malpertuy, Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering, BMC Bioinform. 5 (2004) 114.
[25] S. Oh, D.D. Kang, G.N. Brock, G.C. Tseng, Biological impact of missing-value imputation on downstream analyses of gene expression profiles, Bioinformatics 27 (2011) 78–86.
[26] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. 98 (2001) 5116–5121.
[27] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, Mol. Biol. Cell 9 (1998) 3273–3297.
[28] D.T. Ross, U. Scherf, M.B. Eisen, C.M. Perou, C. Rees, P. Spellman, V. Iyer, et al., Systematic variation in gene expression patterns in human cancer cell lines, Nat. Genet. 24 (2000) 227–235.
[29] C.G. Artieri, R.S. Singh, Molecular evidence for increased regulatory conservation during metamorphosis, and against deleterious cascading effects of hybrid breakdown in Drosophila, BMC Biol. 8 (2010) 26.
[30] R. Pukkila-Worley, F.M. Ausubel, E. Mylonakis, Candida albicans infection of Caenorhabditis elegans induces antifungal immune defenses, PLoS Pathog. 7 (2011) e1002074.
[31] Y.H. Lee, J.B. Andersen, H.T. Song, A.D. Judge, D. Seo, T. Ishikawa, H.G. Woo, Definition of ubiquitination modulator COP1 as a novel therapeutic target in human hepatocellular carcinoma, Cancer Res. 70 (2010) 8264–8269.
[32] E. Rintala, P. Jouhten, M. Toivari, M.G. Wiebe, H. Maaheimo, M. Penttilä, L. Ruohonen, Transcriptional responses of Saccharomyces cerevisiae to shift from respiratory and respirofermentative to fully fermentative metabolism, J. Integr. Plant Biol. 15 (2011) 461–476.
[33] K.Y. Kim, B.J. Kim, G.S. Yi, Reuse of imputed data in microarray analysis increases imputation efficiency, BMC Bioinform. 5 (2004) 160.
[34] L.P. Brás, J.C. Menezes, Improving cluster-based missing value estimation of DNA microarray data, Biomol. Eng. 24 (2007) 273–282.
[35] H. Wang, C.C. Chiu, Y.C. Wu, W.S. Wu, Shrinkage regression-based methods for microarray missing value imputation, BMC Syst. Biol. 7 (2013) s11.
[36] X. Zhang, X. Song, H. Wang, H. Zhang, Sequential local least squares imputation estimating missing value of microarray data, Comput. Biol. Med. 38 (2008) 1112–1120.
[37] Z. Cai, M. Heydari, G. Lin, Iterated local least squares microarray missing value imputation, J. Bioinform. Comput. Biol. 4 (2006) 935–957.
[38] M. Celton, A. Malpertuy, G. Lelandais, A.G. Brevern, Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments, BMC Genom. 11 (2010) 15.

Ye Chen was born in Henan Province, China in 1991. He is now a graduate student with School of Computer and Information, Hefei University of Technology. His research interests include bioinformatics and data mining.

Aiguo Wang was born in Anhui Province, China in 1986. He received his Ph.D. degree at Hefei University of Technology in 2015. He is now a post doctor with School of Computer and Information, Hefei University of Technology. His research interests include data mining, bioinformatics, and activity recognition.

Huitong Ding was born in Shandong Province, China in 1992. He received the B.Sc at Hefei University of Technology in 2015. He has been taking successive postgraduate and doctoral programs of study for pursing his Ph.D. degree since September 2015 with the School of Computer and Information, Hefei University of Technology. His research interests include machine learning and bioinformatics.

Xia Que was born in Guangxi Province, China in 1978. She received her B.Sc and M. Sc at Hefei University of Technology in 2000 and 2006, respectively. She is now a Ph.D. candidate with School of Computer and Information, Hefei University of Technology. Her research interests include data mining and machine learning.

Yabo Li was born in Gansu Province, China in 1989. She received her B.Sc at Lanzhou University in 2011. Now she is a Ph.D. candidate with College of Life Sciences, Lanzhou University. Her research interests include data mining and bioinformatics.

Ning An was born in Gansu Province, China in 1971. He received his B.Sc and M.Sc at Lanzhou University in 1993 and 1996, respectively, and Ph.D. at Pennsylvania State University in 2002. He is now a professor with School of Computer and Information, Hefei University of Technology. His research interests include gerontechnology, healthcare informatics, and spatial information management.

Lili Jiang was born in Heilongjiang Province, China in 1983. She received her doctoral degree at School of Information Science and Engineering, Lanzhou University in 2012. She is now an assistant professor at Umeå University. Her research interests include data mining, natural language processing, and information retrieval.