

SOFTWARE

Open Access



# MVIAeval: a web tool for comprehensively evaluating the performance of a new missing value imputation algorithm

Wei-Sheng Wu\* and Meng-Jhun Jhou

## Abstract

**Background:** Missing value imputation is important for microarray data analyses because microarray data with missing values would significantly degrade the performance of the downstream analyses. Although many microarray missing value imputation algorithms have been developed, an objective and comprehensive performance comparison framework is still lacking. To solve this problem, we previously proposed a framework which can perform a comprehensive performance comparison of different existing algorithms. Also the performance of a new algorithm can be evaluated by our performance comparison framework. However, constructing our framework is not an easy task for the interested researchers. To save researchers' time and efforts, here we present an easy-to-use web tool named MVIAeval (Missing Value Imputation Algorithm evaluator) which implements our performance comparison framework.

**Results:** MVIAeval provides a user-friendly interface allowing users to upload the R code of their new algorithm and select (i) the test datasets among 20 benchmark microarray (time series and non-time series) datasets, (ii) the compared algorithms among 12 existing algorithms, (iii) the performance indices from three existing ones, (iv) the comprehensive performance scores from two possible choices, and (v) the number of simulation runs. The comprehensive performance comparison results are then generated and shown as both figures and tables.

**Conclusions:** MVIAeval is a useful tool for researchers to easily conduct a comprehensive and objective performance evaluation of their newly developed missing value imputation algorithm for microarray data or any data which can be represented as a matrix form (e.g. NGS data or proteomics data). Thus, MVIAeval will greatly expedite the progress in the research of missing value imputation algorithms.

**Keywords:** Web tool, Missing value imputation, Microarray data, Performance index, Performance comparison, Algorithm

## Background

Microarray technology is one of the most powerful high-throughput tools in biomedical and biological research. It has been successfully applied to various studies such as cancer classification [1], drug discovery [2], stress response [3, 4], and cell cycle regulation [5, 6]. Microarray data contain missing values due to various technological limitations such as poor hybridization, spotting problems, insufficient resolution, and fabrication errors. Unfortunately, the

missing values in microarray data would significantly degrade the performance of downstream analyses such as gene clustering and identification of differentially expressed genes [7–9]. Therefore, missing value imputation has become an important pre-processing step in microarray data analyses.

One way to deal with the missing values is to repeat the experiments but it is expensive and time consuming. Another way is to discard the genes with missing values but this loses valuable information. Filling missing values with zeros or with the row average is a simple imputation strategy, but it is far from optimal. Therefore, many advanced algorithms have been developed to impute the missing values in

\* Correspondence: wessonwu@mail.ncku.edu.tw  
Department of Electrical Engineering, National Cheng Kung University,  
Tainan, Taiwan

microarray data [10–12]. The existing algorithms can be divided into four categories [11]: global approach, local approach, hybrid approach and knowledge-assisted approach. Global approach algorithms include SVD [13] and BPCA [14]. Local approach algorithms include KNN [13], SKNN [15], IKNN [16], LS [17], LLS [18], SLLS [19], ILLS [20], Shrinkage LLS [21] and so on. Hybrid approach algorithms include LinCmb [22] and RMI [23]. Knowledge-assisted approach algorithms include GOimpute [24], POCsImpute [25] and HAIimpute [26].

In order to know which algorithm performs best among the dozens of existing ones, an objective and comprehensive performance comparison framework is urgently needed. To meet the need, we previously developed a performance comparison framework [12] which provides 13 testing microarray datasets, three types of performance indices, 9 existing algorithms, and 110 runs of simulation. We found that no single algorithm can perform best for all types of microarray data. The best algorithms are different for different microarray data types (time series and non-time series) and different performance indices, showing the usefulness of our framework for conducting a comprehensive performance comparison [12].

Actually, the most important value of our framework is to give an objective and comprehensive performance evaluation of a new algorithm. Using our framework, bioinformaticians who design new algorithms can easily know their algorithms' performance and then refine their algorithms if needed. However, constructing our framework is not an easy task for the interested bioinformaticians. It involves collecting and processing many microarray raw data from the public domain and using programming languages to implement many existing algorithms and three performance indices. In order to save bioinformaticians' efforts and time, we present an easy-to-use web tool named MVIAeval (Missing Value Imputation Algorithm evaluator) which implements our performance comparison framework.

## Implementation

### Twenty benchmark microarray datasets and twelve existing algorithms used for performance comparison

In MVIAeval, we collected 20 benchmark microarray datasets [27–46] of different species and different types (see Table 1 for details). In addition, we implemented 12 existing algorithms including two global approach algorithms and 10 local approach algorithms (see Table 2 for details). Do note that we did not include hybrid approach algorithms and knowledge-assisted algorithms

because they either are difficult to implement or need extra information from outside data sources which are not always available.

### Three existing performance indices used for performance evaluation

In MVIAeval, we used three existing performance indices for performance evaluation. First, the inverse of the normalized root mean square error (1/NRMSE) [13] is used to measure the numerical similarity between the imputed matrix (generated by an imputation algorithm) and the original complete matrix. Therefore, the higher the 1/NRMSE value is, the better the performance of an imputation algorithm is. Second, the cluster pair proportion (CPP) [47] is used to measure the similarity of the gene clustering results of the imputed matrix and the complete matrix. High CPP value means that the imputed matrix (generated by an imputation algorithm) has very similar gene clustering results as the complete matrix does. Therefore, the higher the CPP value is, the better the performance of an imputation algorithm is. Third, the biomarker list concordance index (BLCI) [7] is used to measure the similarity of the differentially expressed genes identification results of the imputed matrix and the complete matrix. High BLCI value means that differentially expressed genes identified using the imputed matrix (generated by an imputation algorithm) are very similar to those identified using the complete matrix. Therefore, the higher the BLCI value is, the better the performance of an imputation algorithm is. In summary, 1/NRMSE measures the numerical similarity, while CPP and BLCI measure the similarity of downstream analysis results (gene clustering and differentially expressed genes identification) of the imputed matrix and the complete matrix. Fig. 1 shows how the scores of these three performance indices are calculated.

### Evaluating the performance of an algorithm for a benchmark microarray data matrix using a specific performance index

The simulation procedure for evaluating the performance of an imputation algorithm (e.g. KNN) for a given complete benchmark microarray data matrix using a performance index (e.g. CPP) is divided into four steps. **Step 1:** generate five testing matrices having missing values (generated as missing completely at random) with different percentages (1%, 3%, 5%, 8% and 10%) from the complete matrix. **Step 2:** generate five imputed matrices by imputing the missing values in the five testing matrices using KNN. **Step 3:** calculate five CPP scores using the complete matrix

**Table 1** The 20 benchmark microarray datasets of different types and different species

GEO Dataset	Size	Type	Organism	Title
GDS3323[27]	45101x6	Non-time series	<i>Mus musculus</i>	Na <sup>+</sup> /H <sup>+</sup> exchanger 3 deficiency effect on the colon
GDS3215[28]	12625x6	Non-time series	<i>Homo sapiens</i>	13-cis retinoic acid effect on SEB-1 sebocyte cell line
GDS3485[29]	45011x6	Non-time series	<i>Mus musculus</i>	Zinc transporter SLC39A13 deficiency effect on chondrocytes
GDS3476[30]	45011x6	Non-time series	<i>Mus musculus</i>	NF-E2-related factor 2 Nrf2 activation effect on the liver
GDS3197[31]	45101x6	Non-time series	<i>Mus musculus</i>	Transcriptional coactivator PGC-1beta hypomorphic mutation effect on the liver
GDS3149[32]	45101x6	Non-time series	<i>Mus musculus</i>	Suppressor of cytokine signaling 3 deficiency effect on the regenerating liver
GDS2107[33]	15923x6	Non-time series	<i>Rattus norvegicus</i>	Long-term ethanol consumption effect on pancreas
GDS3464[34]	15617x6	Non-time series	<i>Danio rerio</i>	SPT5 mutant embryos
GDS3426[35]	23015x6	Non-time series	<i>Staphylococcus epidermidis</i>	Staphylococcus epidermidis SarZ mutant
GDS3421[36]	10208x6	Non-time series	<i>Escherichia coli</i>	Frag1 cells response to ionic and non-ionic hyperosmotic stress
GDS3360[37]	22575x8	Time series	<i>Homo sapiens</i>	Chlamydia pneumoniae infection effect on HL epithelial cells: time course
GDS2863[38]	31099x6	Time series	<i>Rattus norvegicus</i>	Tienilic acid effect on the liver: time course
GDS5057[39]	34760x8	Time series	<i>Mus musculus</i>	Mepenzolate bromide effect on lung: time course
GDS5055[40]	45307x10	Time series	<i>Mus musculus</i>	Histone demethylase KDM1A deficiency effect on 3 T3-L1 preadipocytes: time course
GDS3428[41]	22283x9	Time series	<i>Homo sapiens</i>	Immature dendritic cell response to butanol fraction of Echinacea purpurea: time course
GDS4484[42]	45101x8	Time series	<i>Mus musculus</i>	Cerebellar neuronal cell response to thyroid hormone: time course
GDS3785[43]	17589x8	Time series	<i>Homo sapiens</i>	Osteoarthritic chondrocytes and healthy mesenchymal stem cell during chondrogenic differentiation: time course
GDS3930[44]	8799x9	Time series	<i>Rattus norvegicus</i>	Bone morphogenic protein effect on cultured sympathetic neurons: time course
GDS4321[45]	10208x8	Time series	<i>Escherichia coli</i>	<i>Escherichia coli</i> O157:H7 response to cinnamaldehyde: time course
GDS3032[46]	22277x8	Time series	<i>Homo sapiens</i>	Quercetin effect on intestinal cell differentiation in vitro: time course

and five imputed matrices. **Step 4:** repeat Steps 1–3 for  $B$  times, where  $B$  is the number of simulation runs per missing percentage. Then the final CPP score of KNN for the given benchmark microarray data matrix is defined as the average of the  $5*B$  CPP scores. Fig. 2 illustrates the whole simulation procedure.

#### Two existing comprehensive performance scores

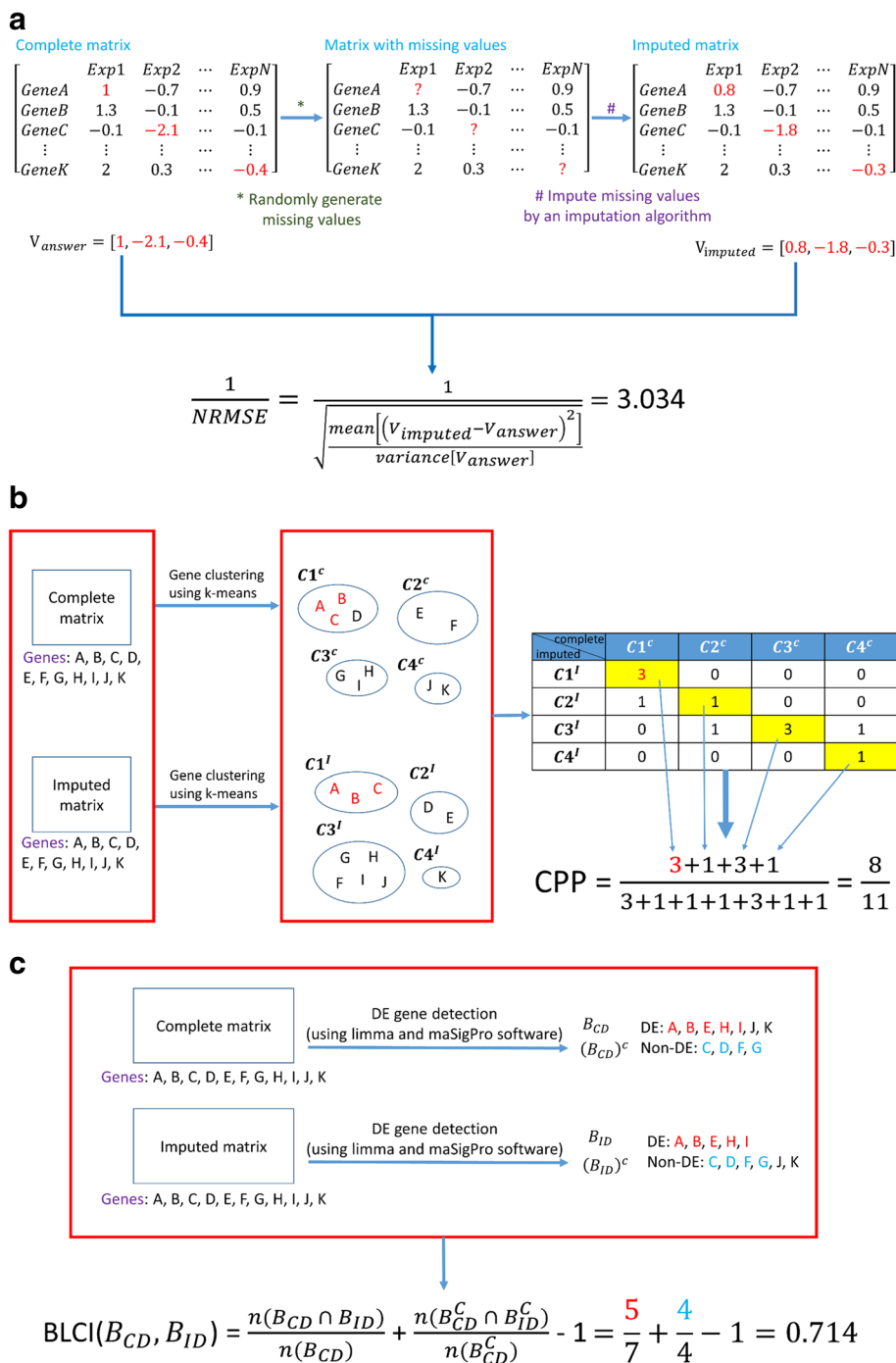
In MVIAeval, we implemented two existing comprehensive performance scores [48, 49] to provide the overall performance comparison results for the selected benchmark microarray datasets and performance indices. The first one, termed the overall ranking score (ORS), is defined as the sum of the rankings of an algorithm for the selected performance indices and benchmark microarray datasets [48, 49]. The ranking of an algorithm for a specific performance index and a specific benchmark microarray dataset is  $d$  if its performance ranks  $\#d$  among all the compared algorithms. For instance, the best algorithm has ranking 1. Therefore, small ORS

indicates that an algorithm has good overall performance.

The other comprehensive performance score, termed the overall normalized score (ONS), is calculated by the sum of the normalized scores for the benchmark

**Table 2** The 12 existing algorithms implemented in MVIAeval

Algorithm	Year of Publication	Category	Reference
SVD	2001	Global	[13]
BPCA	2003	Global	[14]
KNN	2001	Local	[13]
SKNN	2004	Local	[15]
IKNN	2007	Local	[16]
LS	2004	Local	[17]
LLS	2005	Local	[18]
ILLS	2006	Local	[19]
SLLS	2008	Local	[20]
Shrinkage LLS	2013	Local	[21]
Shrinkage SLLS	2013	Local	[21]
Shrinkage ILLS	2013	Local	[21]

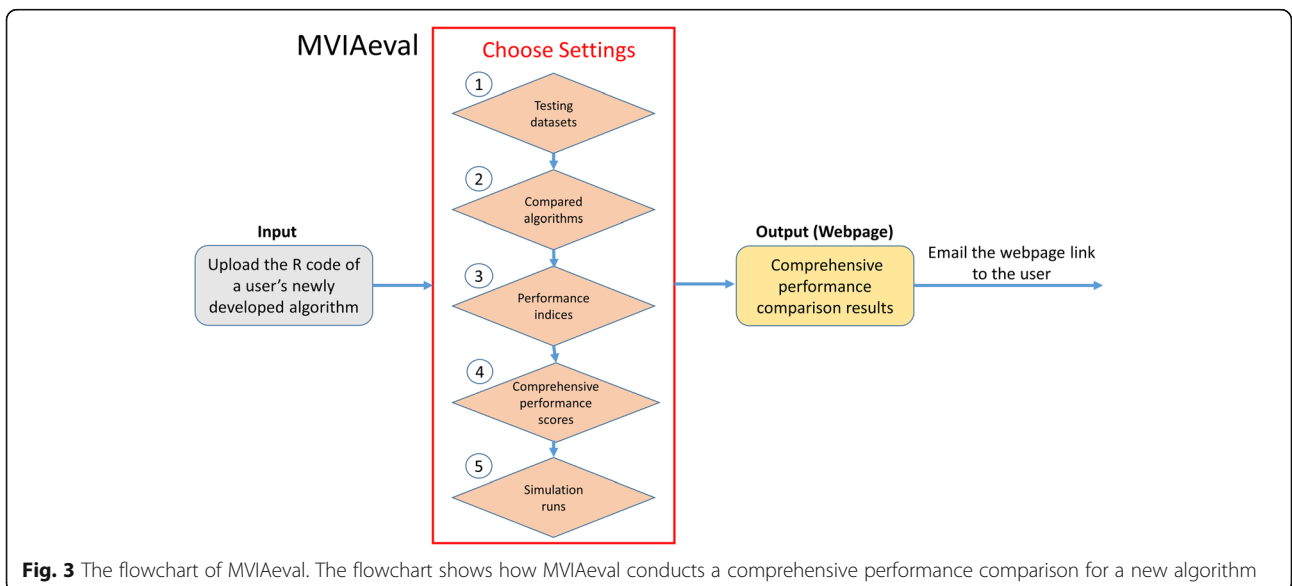
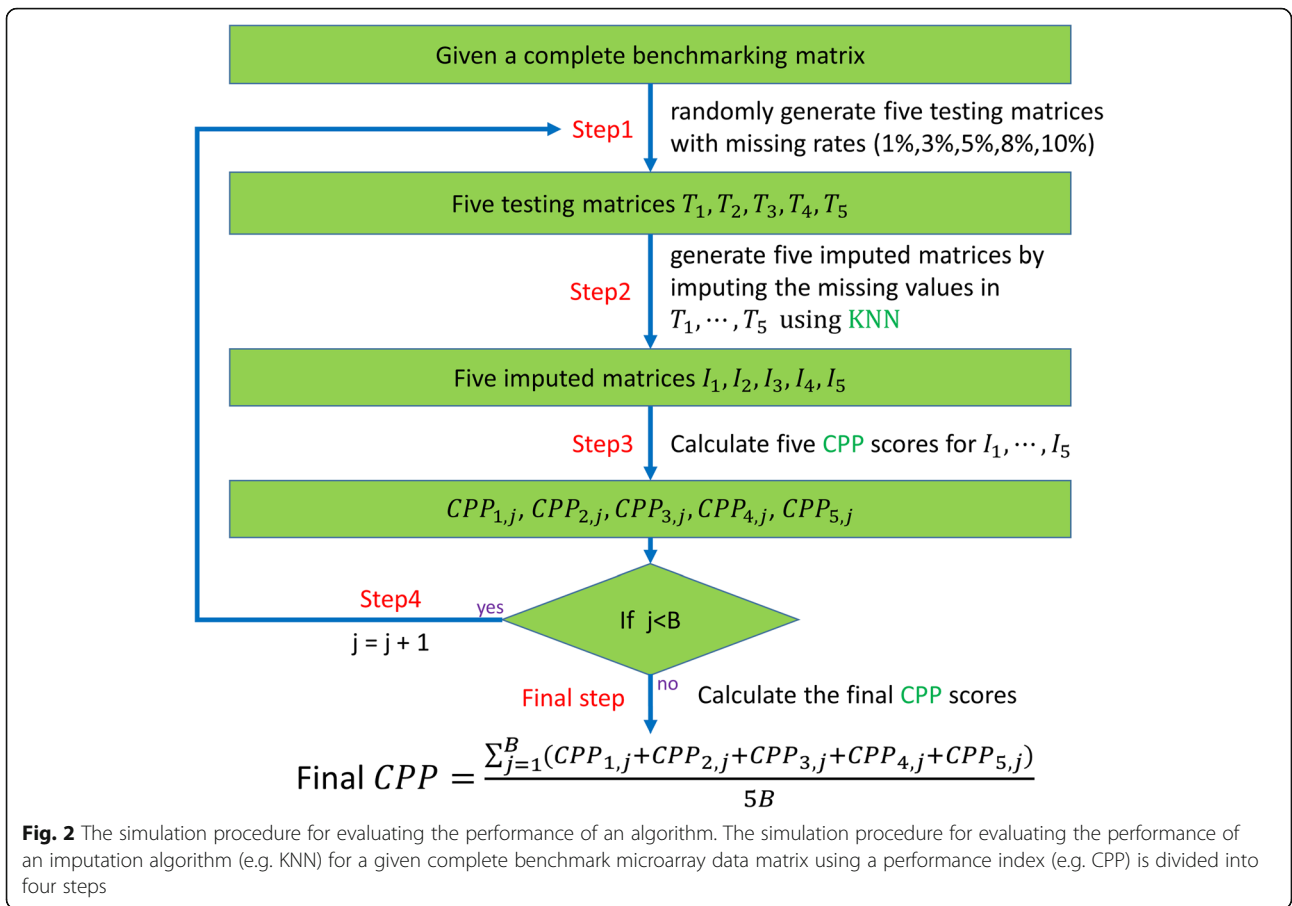


**Fig. 1** Three performance indices implemented in MVIAeval. MVIAeval implements three performance indices, which are **a** 1/NRMSE, **b** CPP and **c** BLCI. Here we provide an example to show how the scores of these three performance indices are calculated

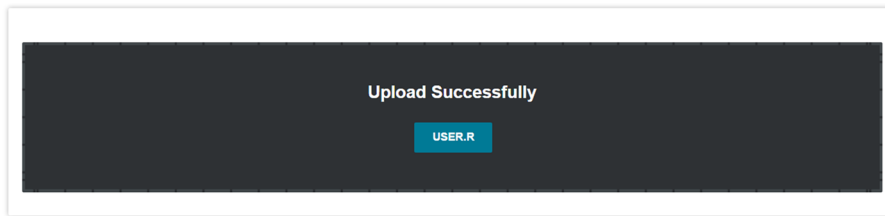
microarray datasets and performance indices [48, 49]. The ONS of the algorithm  $k$  is calculated like the following:

$$\text{ONS}(k) = \sum_{i=1}^I \sum_{j=1}^J N_{ij}(k) = \sum_{i=1}^I \sum_{j=1}^J \left( \frac{S_{ij}(k)}{\max(S_{ij}(1), S_{ij}(2), \dots, S_{ij}(m))} \right)$$

where  $S_{ij}(k)$  and  $N_{ij}(k)$  is the original score and the normalized score of the algorithm  $k$  for the selected performance index  $i$  and benchmark microarray dataset  $j$ , respectively;  $I$  is the number of the selected indices;  $J$  is the number of the selected benchmark microarray datasets and  $m$  is the number of the algorithms being



**a**  
Evaluate the performance of the user's algorithm



Input your e-mail address:

Valid

**b**

<input type="checkbox"/> Check All	Dataset	Dim	Type	Organism	Title
<input type="checkbox"/>	GDS3421	10208*6	Non-time Series	Escherichia coli	Frag1 cells response to ionic and non-ionic hypersmotic stress
<input type="checkbox"/>	GDS4321	10208*8	Time Series	Escherichia coli	Escherichia coli O157:H7 response to cinnamaldehyde: time course
<input checked="" type="checkbox"/>	GDS3215	12625*6	Non-time Series	Homo sapiens	13-cis retinoic acid effect on SEB-1 sebocyte cell line
<input type="checkbox"/>	GDS3464	15617*6	Non-time Series	Danio rerio	SPTS mutant embryos
<input type="checkbox"/>	GDS2107	15923*6	Non-time Series	Rattus norvegicus	Long-term ethanol consumption effect on pancreas
<input checked="" type="checkbox"/>	GDS3785	17589*8	Time Series	Homo sapiens	Osteoarthritic chondrocytes and healthy mesenchymal stem cell during chondrogenic differentiation: time course
<input type="checkbox"/>	GDS3937	22277*6	Time Series	Homo sapiens	Osteogenic effect on intestinal cell differentiation by stem cells: time course

**c**

<input checked="" type="checkbox"/> Uncheck All	Algorithm	Parameters		
		K Nearest Neighbors	Similarity Measure	Number of Iterations
<input checked="" type="checkbox"/>	SVD			
<input checked="" type="checkbox"/>	BPCA			
<input checked="" type="checkbox"/>	KNN	15 (5 < x < 300)	<input type="radio"/> 1/EuDist <input type="radio"/> PCor <input type="radio"/> Angle	
<input checked="" type="checkbox"/>	SKNN	15 (5 < x < 300)	<input type="radio"/> 1/EuDist <input type="radio"/> PCor <input type="radio"/> Angle	
<input checked="" type="checkbox"/>	IKNN	15 (5 < x < 300)	<input type="radio"/> 1/EuDist <input type="radio"/> PCor <input type="radio"/> Angle	5
<input checked="" type="checkbox"/>	LS	300 (5 < x < 300)	<input type="radio"/> 1/EuDist <input type="radio"/> PCor <input type="radio"/> Angle	
<input checked="" type="checkbox"/>	LLS	300 (5 < x < 300)	<input type="radio"/> 1/EuDist <input type="radio"/> PCor <input type="radio"/> Angle	
<input checked="" type="checkbox"/>	SLLS	300 (5 < x < 300)	<input type="radio"/> 1/EuDist <input type="radio"/> PCor <input type="radio"/> Angle	
<input checked="" type="checkbox"/>	ILLS	300 (5 < x < 300)	<input type="radio"/> 1/EuDist <input type="radio"/> PCor <input type="radio"/> Angle	5
<input checked="" type="checkbox"/>	ShrLLS	300 (5 < x < 300)	<input type="radio"/> 1/EuDist <input type="radio"/> PCor <input type="radio"/> Angle	
<input checked="" type="checkbox"/>	ShrSLLS	300 (5 < x < 300)	<input type="radio"/> 1/EuDist <input type="radio"/> PCor <input type="radio"/> Angle	
<input checked="" type="checkbox"/>	IShrLLS	300 (5 < x < 300)	<input type="radio"/> 1/EuDist <input type="radio"/> PCor <input type="radio"/> Angle	5

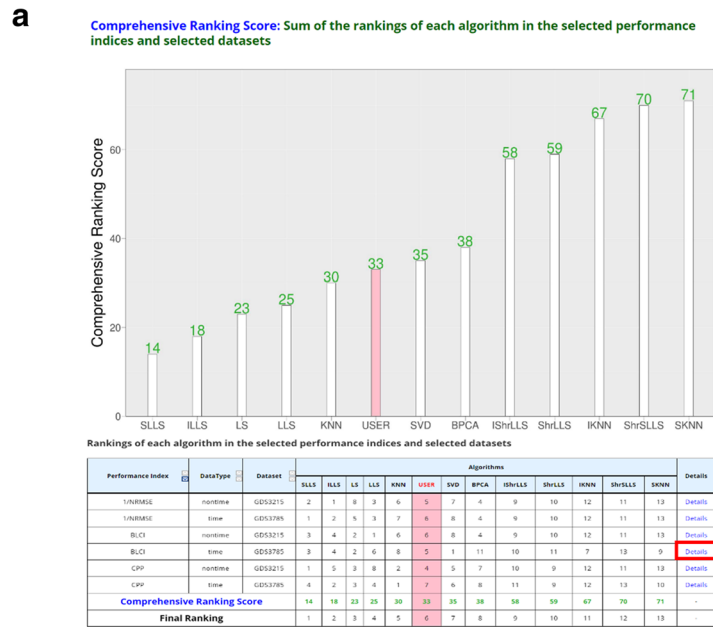
**d**

Performance Indices		Comprehensive Performance Scores	
<input checked="" type="checkbox"/>	1/NRMSE	<input checked="" type="checkbox"/>	Comprehensive Ranking Score
<input checked="" type="checkbox"/>	CPP	<input type="checkbox"/>	Comprehensive Normalized Score
<input checked="" type="checkbox"/>	BLCI		

Simulation Runs



**Fig. 4** The input and five settings of MVAeval. Users need to **a** upload the R code of their new algorithm, **b** select the test datasets among 20 benchmark microarray (time series or non-time series) datasets, **c** select the compared algorithms among 12 existing algorithms, **d** select the performance indices from three existing ones, the comprehensive performance scores from two possible choices, and the number of simulation runs



**\* Details of BLCI Score**

Sim i = ith Simulation	Algorithm												
	SVD	LS	SLLS	ILLS	USER	LLS	IKNN	KNN	SKNN	IShrLLS	ShrLLS	BPCA	ShrSLLS
<b>Average</b>	<b>0.832</b>	<b>0.785</b>	<b>0.777</b>	<b>0.761</b>	<b>0.753</b>	<b>0.747</b>	<b>0.743</b>	<b>0.741</b>	<b>0.729</b>	<b>0.723</b>	<b>0.713</b>	<b>0.713</b>	<b>0.688</b>
Sim1 (Using a matrix with 1% missing values)	0.98	0.99	0.99	0.99	0.99	0.93	0.929	0.797	0.826	0.949	0.94	0.94	0.99
Sim2 (Using a matrix with 1% missing values)	0.966	0.952	0.911	0.911	0.952	0.911	0.948	0.952	0.955	0.937	0.936	0.911	0.94
Sim3 (Using a matrix with 1% missing values)	0.806	0.91	0.911	0.911	0.816	0.816	0.949	0.816	0.933	0.969	0.969	0.911	0.97
Sim4 (Using a matrix with 1% missing values)	0.83	0.816	0.911	0.911	0.761	0.798	0.797	0.791	0.923	0.919	0.919	0.896	0.918
Sim5 (Using a matrix with 1% missing values)	0.93	0.826	0.817	0.791	0.826	0.826	0.945	0.826	0.938	0.871	0.87	0.826	0.926
Final score = average of the above values	0.907	0.893	0.902	0.900	0.899	0.893	0.891	0.886	0.891	0.890	0.877	0.877	0.868

**Fig. 5** The output of MVIaEval. For demonstration purpose, we upload the R code of a sample algorithm as the user's new algorithm and select two benchmark datasets (GDS3215 and GDS3785), 12 existing algorithms, three performance indices, the overall ranking score as the comprehensive performance score, and 25 simulation runs. **a** The webpage of the comprehensive performance comparison results shows that the overall performance of the user's algorithm (denoted as USER) ranks six among all the 13 compared algorithms. **b** By clicking "details" in the row of BLCI for the benchmark dataset GDS3785, users can see the performance comparison results using only BLCI score for the benchmark dataset GDS3785. It can be seen that the user's algorithm ranks five among the 13 compared algorithms using only BLCI score for the benchmark dataset GDS3785. The details of BLCI score for each algorithm can also be found



compared. Note that  $0 \leq N_{ij}(k) \leq 1$  and  $N_{ij}(k) = 1$  when the algorithm  $k$  performs best for the selected performance index  $i$  and benchmark microarray dataset  $j$  (i.e.  $S_{ij}(k) = \max(S_{ij}(1), S_{ij}(2), \dots, S_{ij}(m))$ ). Therefore, large ONS indicates that an algorithm has good overall performance.

## Results and discussion

### Usage

Figure 3 illustrates the usage of MVIAeval. The easy-to-use web interface allows users to upload the R code of their newly developed algorithm. Subsequently five types of settings of MVIAeval need to be set. First, the test datasets have to be chosen from 20 benchmark microarray datasets. The collected benchmark datasets consist of two types of data: 10 non-time series data and 10 time series data. Second, the compared algorithms have to be chosen from 12 existing algorithms. The collected existing algorithms consist of two global approach algorithms and 10 local approach algorithms. Third, the performance indices have to be chosen from three existing ones (1/NRMSE, CPP and BLCI). Fourth, the comprehensive performance scores have to be chosen from two existing ones (ORS and ONS). Fifth, the number of simulation runs have to be specified. The larger the number of simulation runs is, the more accurate the comprehensive performance comparison result is. But be cautious that the simulation time increases linearly with the number of simulation runs. After submission, a comprehensive performance comparison between the user's algorithm and the selected existing algorithms is executed by MVIAeval using the selected benchmark datasets and performance indices. Then a webpage of the comprehensive performance comparison results is generated and the webpage link is sent to the users by e-mails.

### A case study

In MVIAeval, the R code of a sample algorithm is provided. For demonstration purpose, we regard the sample algorithm as the user's newly developed algorithm and would like to use MVIAeval to conduct a comprehensive performance comparison of this new algorithm (denoted as USER) to various existing algorithms. For example, users may upload the R code of the new algorithm and select (i) two benchmark datasets, (ii) 12 existing algorithms, (iii) three performance indices, (iv) the overall ranking score as the comprehensive performance score, and (v) 25 simulation runs (see Fig. 4). After submission, MVIAeval outputs the comprehensive comparison results in both tables and figures. Among the 13 compared algorithms, the overall performance of the new algorithm ranks six (see Fig. 5). Actually, MVIAeval can provide the performance comparison results in many scenarios

(see Table 3). It can be concluded that the new algorithm is mediocre because its performance is always in the middle of all the 13 compared algorithms in different data types (time series or non-time series), different performance indices (1/NRMSE, BLCI or CPP) and different comprehensive performance scores (ORS or ONS). Receiving the comprehensive comparison results from MVIAeval, researchers immediately know that there is much room to improve the performance of their new algorithm.

## Conclusions

Missing value imputation is an inevitable pre-processing step of microarray data analyses. This is why the computational imputation of the missing values in microarray data has become a hot research topic. The newest algorithm is published in year 2016 [50] and we believe that many new algorithms will be developed in the near future. Using MVIAeval, bioinformaticians can easily get a comprehensive and objective performance comparison results of their new algorithm. Therefore, bioinformaticians now can focus on developing new algorithms instead of putting a lot of efforts for conducting a comprehensive and objective performance evaluation of their new algorithm. In conclusion, MVIAeval will definitely be a very useful tool for developing missing value imputation algorithms.

**Table 3** MVIAeval can provide the performance comparison results in many scenarios

Performance Index	Benchmark datasets	Ranking of USER using ORS	Ranking of USER using ONS
1/NRMSE	Five Time Series [37–41]	5	6
	Five Non-time Series [27–31]	6	6
CPP	Five Time Series [37–41]	7	9
	Five Non-time Series [27–31]	11	8
BLCI	Five Time Series [37–41]	3	4
	Five Non-time Series [27–31]	7	7
1/NRMSE + CPP + BLCI	Five Time Series [37–41]	6	7
	Five Non-time Series [27–31]	6	6

The performance comparison results of the user's algorithm (denoted as USER) and various existing algorithms using different types of datasets (time series or non-time series), different performance indices (1/NRMSE, CPP or BLCI), and different overall performance scores (overall ranking score (ORS) or overall normalized score (ONS)) are shown. More details could be seen at [http://cos.bi.ee.ncku.edu.tw/MVIAeval/A\\_Case\\_Study](http://cos.bi.ee.ncku.edu.tw/MVIAeval/A_Case_Study)



**Abbreviations**

MVIAeval: Missing value imputation algorithm evaluator; NRMSE: Normalized root mean square error; CPP: Cluster pair proportion; BLCI: Biomarker list concordance index; ORS: Overall ranking score; ONS: Overall normalized score

**Acknowledgements**

The authors thank Dr. Jagat Rathod and Dr. Fu-Jou Lai for proofreading this manuscript.

**Funding**

The publication of this paper was funded by Ministry of Science and Technology of Taiwan MOST-105-2221-E-006-203-MY2 and MOST-103-2221-E-006 -174 -MY2.

**Data and material availability**

Project name: MVIAeval  
Project home page: <http://cosbi.ee.ncku.edu.tw/MVIAeval/>  
Operating system(s): platform independent  
Programming language: R, Javascript and PHP  
Other requirements: Internet connection  
License: none required  
Any restrictions to use by non-academics: no restriction

**Authors' contributions**

WSW conceived the research topic, designed the website structure, provided essential guidance and wrote the manuscript. MJJ collected benchmark microarray datasets, constructed MVIAeval web tool and prepared all the figures. Both authors approved the final manuscript.

**Competing interests**

The authors declare that they have no competing interests.

**Consent for publication**

Not applicable.

**Ethics approval and consent to participate**

Not applicable.

Received: 15 May 2016 Accepted: 15 December 2016

Published online: 13 January 2017

**References**

- Colombo PE, Milanezi F, Weigelt B, Reis-Filho JS. Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction. *Breast Cancer Res.* 2011;13(3):212.
- Wang S, Cheng Q. Microarray analysis in drug discovery and clinical applications. *Methods Mol Biol.* 2006;316:49–65.
- Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell.* 2000;11(12):4241–57.
- Wu WS, Li WH. Identifying gene regulatory modules of heat shock response in yeast. *BMC Genomics.* 2008;9:439.
- Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell.* 1998;9(12):3273–97.
- Wu WS, Li WH. Systematic identification of yeast cell cycle transcription factors using multiple data sources. *BMC Bioinformatics.* 2008;9:522.
- Oh S, Kang DD, Brock GN, Tseng GC. Biological impact of missing-value imputation on downstream analyses of gene expression profiles. *Bioinformatics.* 2011;27(1):78–86.
- Tuikkala J, Elo LL, Nevalainen OS, Aittokallio T. Missing value imputation improves clustering and interpretation of gene expression microarray data. *BMC Bioinformatics.* 2008;9:202.
- Scheel I, Aldrin M, Glad IK, Sørum R, Lyng H, Frigessi A. The influence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics.* 2005;21(23):4272–9.
- Aittokallio T. Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Brief Bioinform.* 2010;11(2):253–64.
- Liew AW, Law NF, Yan H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Brief Bioinform.* 2011;12(5):498–513.
- Chiu CC, Chan SY, Wang CC, Wu WS. Missing value imputation for microarray data: a comprehensive comparison study and a web tool. *BMC Syst Biol.* 2013;7 Suppl 6:S12.
- Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. *Bioinformatics.* 2001;17(6):520–5.
- Oba S, Sato MA, Takemasa I, Monden M, Matsubara K, Ishii S. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics.* 2003;19(16):2088–96.
- Kim KY, Kim BJ, Yi GS. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinformatics.* 2004;5:160.
- Brás LP, Menezes JC. Improving cluster-based missing value estimation of DNA microarray data. *Biomol Eng.* 2007;24(2):273–82.
- Bø TH, Dysvik B, Jonassen I. LSImpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Res.* 2004;32(3):e34.
- Kim H, Golub GH, Park H. Missing Value Estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics.* 2005;21(2):187–98.
- Cai Z, Heydari M, Lin G. Iterated local least squares microarray missing value imputation. *J Bioinform Comput Biol.* 2006;4(5):935–57.
- Zhang X, Song X, Wang H, Zhang H. Sequential local least squares imputation estimating missing value of microarray data. *Comput Biol Med.* 2008;38(10):1112–20.
- Wang H, Chiu CC, Wu YC, Wu WS. Shrinkage regression-based methods for microarray missing value imputation. *BMC Syst Biol.* 2013;7 Suppl 6:S11.
- Jörnsten R, Wang HY, Welsh WJ, Ouyang M. DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics.* 2005;21(22):4155–61.
- Li H, Zhao C, Shao F, Li GZ, Wang X. A hybrid imputation approach for microarray missing value estimation. *BMC Genomics.* 2015;16 Suppl 9:S1.
- Tuikkala J, Elo L, Nevalainen O, Aittokallio T. Improving missing value estimation in microarray data with gene ontology. *Bioinformatics.* 2006;22(5):566–72.
- Gan X, Liew AW, Yan H. Microarray missing data imputation based on a set theoretic framework and biological knowledge. *Nucleic Acids Res.* 2006;34(5):1608–19.
- Xiang Q, Dai X, Deng Y, He C, Wang J, Feng J, Dai Z. Missing value imputation for microarray gene expression data using histone acetylation information. *BMC Bioinformatics.* 2008;9:252.
- Laubitz D, Larmonier CB, Bai A, Midura-Kiela MT, Lipko MA, Thurston RD, Kiela PR, Ghishan FK. Colonic gene expression profile in NHE3-deficient mice: evidence for spontaneous distal colitis. *Am J Physiol Gastrointest Liver Physiol.* 2008;295(1):G63–77.
- Nelson AM, Zhao W, Gilliland KL, Zaenglein AL, Liu W, Thiboutot DM. Neutrophil gelatinase-associated lipocalin mediates 13-cis retinoic acid-induced apoptosis of human sebaceous gland cells. *J Clin Invest.* 2008;118(4):1468–78.
- Fukada T, Civic N, Furuichi T, Shimoda S, Mishima K, Higashiyama H, Idaira Y, Asada Y, Kitamura H, Yamasaki S, Hojyo S, Nakayama M, Ohara O, Koseki H, Dos Santos HG, Bonafe L, Ha-Vinh R, Zankl A, Unger S, Kraenzlin ME, Beckmann JS, Saito I, Rivolta C, Ikegawa S, Superti-Furga A, Hirano T. The zinc transporter SLC39A13/ZIP13 is required for connective tissue development; its involvement in BMP/TGF-beta signaling pathways. *PLoS One.* 2008;3(11):e3642.
- Osburn WO, Yates MS, Dolan PD, Chen S, Liby KT, Sporn MB, Taguchi K, Yamamoto M, Kensler TW. Genetic or pharmacologic amplification of nrf2 signaling inhibits acute inflammatory liver injury in mice. *Toxicol Sci.* 2008;104(1):218–27.
- Vianna CR, Huntgeburth M, Coppari R, Choi CS, Lin J, Krauss S, Barbatelli G, Tzamelis I, Kim YB, Cinti S, Shulman GI, Spiegelman BM, Lowell BB. Hypomorphic mutation of PGC-1beta causes mitochondrial dysfunction and liver insulin resistance. *Cell Metab.* 2006;4(6):453–64.
- Riehle KJ, Campbell JS, McMahan RS, Johnson MM, Beyer RP, Bammler TK, Fausto N. Regulation of liver regeneration and hepatocarcinogenesis by suppressor of cytokine signaling 3. *J Exp Med.* 2008;205(1):91–103.

33. Kubisch CH, Gukovsky I, Lugea A, Pandol SJ, Kuick R, Misek DE, Hanash SM, Logsdon CD. Long-term ethanol consumption alters pancreatic gene expression in rats: a possible connection to pancreatic injury. *Pancreas*. 2006;33(1):68–76.
34. Krishnan K, Salomonis N, Guo S. Identification of Spt5 target genes in zebrafish development reveals its dual activity in vivo. *PLoS One*. 2008;3(11):e3621.
35. Wang L, Li M, Dong D, Bach TH, Sturdevant DE, Vuong C, Otto M, Gao Q. SarZ is a key regulator of biofilm formation and virulence in *Staphylococcus epidermidis*. *J Infect Dis*. 2008;197(9):1254–62.
36. Shabala L, Bowman J, Brown J, Ross T, McMeekin T, Shabala S. Ion transport and osmotic adjustment in *Escherichia coli* in response to ionic and non-ionic osmotic. *Environ Microbiol*. 2009;11(1):137–48.
37. Alvesalo J, Greco D, Leinonen M, Raitila T, Vuorela P, Auvinen P. Microarray analysis of a *Chlamydia pneumoniae*-infected human epithelial cell line by use of gene ontology hierarchy. *J Infect Dis*. 2008;197(1):156–62.
38. Pacitto SR, Uetrecht JP, Boutros PC, Popovic M. Changes in gene expression induced by tienilic Acid and sulfamethoxazole: testing the danger hypothesis. *J Immunotoxicol*. 2007;4(4):253–66.
39. Tanaka K, Ishihara T, Sugizaki T, Kobayashi D, Yamashita Y, Tahara K, Yamakawa N, Iijima K, Mogushi K, Tanaka H, Sato K, Suzuki H, Mizushima T. Mepenzolate bromide displays beneficial effects in a mouse model of chronic obstructive pulmonary disease. *Nat Commun*. 2013;4:2686.
40. Hanzu FA, Musri MM, Sánchez-Herrero A, Claret M, Esteban Y, Kaliman P, Gomis R, Párrizas M. Histone demethylase KDM1A represses inflammatory gene expression in preadipocytes. *Obesity (Silver Spring)*. 2013;21(12):E616–25.
41. Wang CY, Staniforth V, Chiao MT, Hou CC, Wu HM, Yeh KC, Chen CH, Hwang PI, Wen TN, Shyur LF, Yang NS. Genomics and proteomics of immune modulatory effects of a butanol fraction of *echinacea purpurea* in human dendritic cells. *BMC Genomics*. 2008;9:479.
42. Chatonnet F, Guyot R, Picou F, Bondesson M, Flamant F. Genome-wide search reveals the existence of a limited number of thyroid hormone receptor alpha target genes in cerebellar neurons. *PLoS One*. 2012;7(5):e30703.
43. Bernstein P, Sticht C, Jacobi A, Liebers C, Manthey S, Stiehler M. Expression pattern differences between osteoarthritic chondrocytes and mesenchymal stem cells during chondrogenic differentiation. *Osteoarthritis Cartilage*. 2010;18(12):1596–607.
44. Garred MM, Wang MM, Guo X, Harrington CA, Lein PJ. Transcriptional responses of cultured rat sympathetic neurons during BMP-7-induced dendritic growth. *PLoS One*. 2011;6(7):e21754.
45. Visvalingam J, Hernandez-Doria JD, Holley RA. Examination of the genome-wide transcriptional response of *Escherichia coli* O157:H7 to cinnamaldehyde exposure. *Appl Environ Microbiol*. 2013;79(3):942–50.
46. Dihal AA, Tilburgs C, van Erk MJ, Rietjens IM, Woutersen RA, Stierum RH. Pathway and single gene analyses of inhibited Caco-2 differentiation by ascorbate-stabilized quercetin suggest enhancement of cellular processes associated with development of colon cancer. *Mol Nutr Food Res*. 2007; 51(8):1031–45.
47. de Brevern AG, Hazout S, Malpertuy A. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinformatics*. 2004;5:114.
48. Lai FJ, Chang HT, Huang YM, Wu WS. A comprehensive performance evaluation on the prediction results of existing cooperative transcription factors identification algorithms. *BMC Syst Biol*. 2014;8 Suppl 4:S9.
49. Lai FJ, Chang HT, Wu WS. PCTFPeval: a web tool for benchmark newly developed algorithms for predicting cooperative transcription factor pairs in yeast. *BMC Bioinformatics*. 2015;16 Suppl 18:S2.
50. Yang Y, Xu Z, Song D. Missing value imputation for microRNA expression data by using a GO-based similarity measure. *BMC Bioinformatics*. 2016;17 Suppl 1:10.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

