Missing value imputation for gene expression data: computational techniques to recover missing data from available information

Alan Wee-Chung Liew
School of Information and Communication Technology
Gold Coast Campus, Griffith University, QLD4222, Australia
Email: a.liew@griffith.edu.au

Ngai-Fong Law
Centre for Signal Processing, Department of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hung Hom, Hong Kong
Email: ennflaw@polyu.edu.hk

Hong Yan
Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Hong Kong
School of Electrical and Information Engineering, University of Sydney, NSW 2006, Sydney, Australia
Email: h.yan@cityu.edu.hk

Abstract
Microarray gene expression data generally suffers from missing value problem due to a variety of experimental reasons. Since the missing data points can adversely affect downstream analysis, many algorithms have been proposed to impute missing values. In this survey, we provide a comprehensive review of existing missing value imputation algorithms, focusing on their underlying algorithmic techniques and how they utilize local or global information from within the data, or their use of domain knowledge during imputation. In addition, we describe how the imputation results can be validated and the different ways to assess the performance of different imputation algorithms, as well as a discussion on some possible future research directions. It is hoped that this review will give the readers a good understanding of the current development in this field and inspire them to come up with the next generation of imputation algorithms.

Keywords: missing value imputation, gene expression data, gene expression analysis, information recovery

1. Introduction

Microarray technology has been one of the most useful tools in functional genomics research [1]. It has been used widely in numerous studies over a broad range of biological disciplines, such as cancer classification [2], identification of genes relevant to a certain diagnosis or therapy [3], investigation the mechanism of drug action and cancer prognosis [4, 5]. Using this technology, the relative expression levels in two or more mRNA populations can be analyzed for tens of thousands of genes simultaneously.

In a cDNA microarray experiment, two samples of fluorescence labeled cRNA targets which are reversed transcribed from mRNA purified from cellular contents are hybridized onto a cDNA microarray. If the cDNA sequence on the target is complementary to the DNA probe on a given spot, that cDNA will hybridize to the spot, where it will be detected by a laser scanner. The laser scanner will scan at the two specific wavelengths corresponding to the two fluorescence dyes, giving a two-channel signal. The ratio of the two fluorescence intensities at each spot indicates the relative

abundance of the corresponding DNA sequence in the two cDNA samples. As one sample is the reference, the ratio expresses the extent to which the other sample is differentially expressed with respect to the reference. By examining the expression ratio of each spot, gene expression study can be performed.

The gene expression data from microarray experiments is usually in the form of large matrices of expression levels of genes (rows) under different experimental conditions (columns). In spite of the wide spread use of microarray technology, microarray data often suffers from missing value problem. Microarray data can contain up to 10% missing values and in some datasets, up to 90% of genes have one or more missing values [6, 7]. Missing values occur due to a variety of reasons including hybridization failures, artifacts on the microarray, insufficient resolution, image noise and corruption, or they may occur systematically as a result of the spotting process [8]. A spot that has negative background corrected intensity would normally be declared as missing. In addition, suspicious values are often flagged as missing too. The occurrence of missing values in gene expression data can adversely affect downstream analysis. Missing values have been found to give nontrivial negative effect on some popular algorithms, such as hierarchical clustering and support vector machine (SVM) classifier, and many analysis methods such as principal component analysis (PCA) and singular value decomposition (SVD) cannot be applied to data with missing values.

Instead of repeating the experiments, one can attempt to estimate the missing values by imputation. Many algorithms have been proposed for missing value imputation in gene expression data. A quick search in PubMed on the phrase *missing value imputation* in the Title/Abstract field returns 25 articles, in which 15 articles were published during 2008-2010. Early approaches to deal with missing values include discarding the rows containing missing values, replacing missing values by zeros, or imputing missing values by row average or median [9]. Recent research demonstrates that missing values estimation can be significant improved by incorporating information about the data into the imputation.

In this paper, we attempt to provide a comprehensive survey of existing imputation algorithms in terms of their conceptual and algorithmic basis. A recent review in [62] discusses in broad term the missing value problem and existing approaches from the perspective of a biologist or data analyst. In contrast, this paper provides in depth discussion on the algorithmic aspects of missing data estimation in existing methods from the perspective of a computer scientist or engineer, who works on algorithm and software development. We survey existing missing value imputation algorithms in terms of how information is incorporated into the imputation process, how the imputation results can be validated, how their performance can be compared in terms of internal and external measures, as well as factors that could affect their performance. Finally, we conclude the survey with a discussion on some possible future research directions.

2. The missing value problem

In a typical gene expression data matrix, the rows are the genes (or oligonucleotides) under investigation and the columns are the experimental conditions or time points. The gene expression data matrix is obtained by performing a series of microarray experiments on the same set of genes, one for each column. Let the gene expression data be represented as an $M \times N$ matrix Y where the entries of Y are the expression ratios for M genes under N different conditions or time points. Then the element y_{ij} denotes the expression level of the i-th gene in the j-th experiment. The objective of missing value imputation is to estimate the missing entries given the incomplete gene expression data matrix Y.

Missing value imputation involves exploiting information about the data to estimate the missing entries. In general, there are two types of information available. The first type of information is the correlation structure between entries in the data matrix. In gene expression data matrix, correlation between rows exists due to the fact that genes involved in similar cellular processes usually have

similar expression profiles. Similarly, correlation between columns exists since the set of genes is expected to behave similarly under similar conditions. Hence, it is possible to estimate the missing entries based on subset of related genes or subset of related conditions. The second type of information is domain knowledge about the data or the processes that generate the data. The domain knowledge can be used to regularize the estimation such that more plausible estimates are obtained. More recent gene expression missing value imputation algorithms have tried to incorporate information about the underlying biological processes to improve the imputation accuracy. The more one knows about the biological process behind the data, the better one can constrain the solution to the missing value imputation problem.

3. Missing Value Imputation Techniques

Based on the type of information used in the algorithm, we categorize existing algorithms into four different classes: (1) global approach, (2) local approach, (3) hybrid approach, and (4) knowledge assisted approach. The table below shows our grouping of imputation algorithms considered in this paper.

<< Table 1 here >>

3.1 Global approach

Algorithms in this category perform missing value imputation based on global correlation information derived from the entire data matrix. They assume the existence of a global covariance structure among all genes or samples in the expression matrix. When this assumption is not appropriate, i.e., when the genes exhibit dominant local similarity structures, their imputation becomes less accurate. Well known imputation algorithms in this category include SVD imputation (SVDimpute) [9] and Bayesian principal component analysis (BPCA) [10]. In SVDimpute, the singular value decomposition is used to obtain a set of mutually orthogonal expression patterns, called eigengenes, which can be linearly combined to approximate the expression of all genes in the dataset. SVDimpute first regresses the gene against the k most significant eigengenes and then use the coefficients of the regression to reconstruct the missing values from a linear combination of the k eigengenes. In BPCA, the N-dimensional gene expression vectors y is expressed as a linear combination of K principal axis vectors v_l , i.e., $y = \sum_{l=1}^{K} w_l v_l + \varepsilon$, where the factor scores w_l and the residual error ε are regarded as normally distributed random variables in the probabilistic PCA model. An EM-like algorithm is then used to estimate the posterior distributions of the model parameter and the missing values simultaneously. BPCA suggests setting $K = N - 1$, but if the effective dimension of the data set is smaller than K, the algorithm automatically suppresses the redundant principal axes.

3.2 Local approach

In contrast to global approach, algorithms in this category exploit only local similarity structure in the dataset for missing value imputation. Only a subset of genes that exhibits high correlation with the gene containing the missing values is used to compute the missing values in the gene. Some of the earliest and well-known imputation algorithms, such as, K nearest-neighbor imputation (KNNimpute) [9], least square imputation (LSimpute) [11], local least square imputation (LLSimpute) [12], are among this category.

KNNimpute [9] is perhaps one of the earliest and most frequently used missing value imputation algorithms. KNNimpute uses pairwise information between the target gene with missing values and the K nearest reference genes to impute the missing values. The missing value j in the target gene is estimated as the weighted average of the j-th component of the K reference genes with the weights set proportional to the inverse of the Euclidean distance between the target and the reference genes.

KNNimpute performs well when strong local correlation exists between genes in the data. Several modifications to the basic KNNimpute algorithm have been proposed [13, 14]. In sequential KNNimpute (SKNNimpute) [13], imputed genes are reused in later missing value imputation of other genes. In SKNNimpute, the data matrix is first split into two sets, where the first set, i.e. the reference set, consists of genes with no missing values and the second set, i.e. the target set, consists of genes with missing values that are ranked according to the missing rate. The missing values are estimated sequentially using KNNimpute, starting with the gene having the smallest missing rate in the target set. Once all the missing values in a target gene are imputed, the target gene is moved to the reference set to be used for subsequent imputation of the remaining genes in the target set. In contrast to the single pass SKNNimpute, the iterative KNN imputation algorithm (IKNNimpute) of [14] uses an iterative process to refine the missing value estimates. During each iteration, K closest reference genes selected from the previously imputed complete matrix are used to refine the missing values estimated of the target gene. The iteration terminates when the sum of square difference between the current and the previous estimated complete matrix falls below a threshold.

In Gaussian mixture clustering imputation (GMCimpute) [15], the data is clustered into S components Gaussian mixtures using the EM algorithm. Then the S estimates of the missing value, one from each component, are averaged to obtain the final estimate of the missing value. The clustering and estimation steps are iterated until the cluster memberships of two consecutive iterations are identical. GMCimpute uses the local correlation information in the data through the mixture components. Nevertheless, unlike KNNimpute where K is usually taken to be around 10 to 20, the mixture components in GMCimpute can consists of hundreds of genes. Hence, GMCimpute is able to use more global correlation information.

A number of local imputation algorithms use the concept of least square regression to estimate the missing values. In least square imputation (LSimpute) [11], the target gene y and the reference gene x are assumed to be related by the linear regression model $y = \alpha + \beta x + \varepsilon$. LSimpute first select the K most correlated genes based on absolute Pearson correlation values. Then a least square estimate of the missing value is obtained from each of the K selected genes using single (pairwise) regression. Finally, the K estimates are linearly combined to form the final estimate. LSimpute also estimates the missing values by considering the correlation structure between columns and combines the estimates from row-wise (LSimpute_gene) and column-wise (LSimpute_array) imputations linearly to improve the imputation accuracy. Unlike LSimpute, Local least square imputation (LLSimpute) [12] uses a multiple regression model $y = X^T \alpha + \varepsilon$ to impute the missing values from all K reference genes simultaneously. Despite its simplicity, LLSimpute has been shown to be highly competitive compared to KNNimpute and the much more complex BPCA [12, 16].

Several extensions to the basic LLSimpute algorithm have been proposed [17-19]. In sequential LLSimpute (SLLSimpute) [17], the imputation is performed sequentially starting from the gene with the least missing rate, and the imputed genes are then used for later imputation of other genes. However, only genes with missing rate below a certain threshold are reused since genes with many imputed missing values are less reliable. SLLSimpute has been shown to exhibit better performance than LLSimpute due to the reuse of genes with missing values. In iterated LLSimpute (ILLSimpute) [18], different target genes are allowed to have different number of reference genes. The number of reference genes is chosen based on a distance threshold which is proportional to the average distance of all other genes to the target gene. ILLSimpute iteratively refines the imputation by using the imputed results from previous iteration to re-select the set of coherent genes to re-estimate the missing values until a preset number of iterations is reached. ILLSimpute has been shown to outperform the basic LLSimpute, KNNimpute, and BPCA due to these modifications. The robust least square estimation with principal components (RLSP) method [19] uses the principle components of the K reference genes to impute the missing values and show that this helps to resolve the problem of collinearity in LLSimpute.

In the regression with Bayesian gene selection (BGSregress) method of [20], the target gene and the *K* reference genes are assumed to be related by a multiple regression model. However, instead of choosing the reference genes based on a similarity measure, they are selected by using the Bayesian variable selection algorithm in which a Gibbs sampler is used to obtain the *K* reference genes having the strongest predictive power for the target gene. Then, the missing values are estimated by solving a linear regression problem, either by a least square formulation or by using another Gibbs sampler, to estimate the regression coefficients. It was shown that the proposed method significantly outperforms KNNimpute in imputation accuracy in terms of normalized RMS error (NRMSE).

Collateral missing value imputation (CMVE) [21] is a local imputation technique that utilizes the concept of multiple parallel estimation of missing values to improve the final estimation. In CMVE, the *K* reference genes are selected by using the absolute covariance value between the reference gene and the target gene. The first estimate of the missing value is then obtained by solving a single least square (LS) regression problem as in LSimpute. Two more estimates of the missing values are obtained from the reference genes by solving a non-negative least square (NNLS) regression problem. By combining the three estimates, CMVE was able to demonstrate better accuracy in NRMSE over BPCA, KNN, and LSimpute on several data sets involving ovarian cancer and yeast sporulation time series data. Ameliorative Missing Value Imputation (AMVI) [22] improves on the CMVE by using Monte Carlo simulation to determine the optimal number of reference genes *K* that minimizes the NRMSE score. Experimental results in [22] indicated that significant improvement in imputation accuracy can be achieved over CMVE with the optimal *K* determined this way.

It is well known that time series expression profile exhibits strong dependency between observations. This fact has been used in deriving a continuous representation of time series gene expression profiles using cubic splines [24] and in the reconstruction of unevenly sampled profiles using B-spline bases [25]. Although the two methods are not posed as missing value imputation methods, they nevertheless allow the prediction of missing entries in the time series profiles. Recently, an imputation method that exploits correlation between genes as well as between time points in the gene expression profile was proposed [26]. The correlation between time points in a profile is captured by an autoregressive (AR) model that generates the profile. The correlation between genes is captured by assuming the highly correlated genes are generated by the same AR process. The algorithm, called autoregressive least square imputation (ARLSimpute), first select a set of *K* most correlated genes based on Euclidean distance. Then the *K* genes are used to jointly estimate the AR coefficients and the missing values are estimated by solving a least square problem using the estimated AR coefficients. The experimental results in [26] shown that the imputation accuracy of ARLSimpute in terms of NRMSE for time series expression data is significantly better than imputation algorithms that ignore the within profile correlation.

3.3 Hybrid approach

The correlation structure in the data affects the performance of imputation algorithms. If the dataset is heterogeneous, local correlation between genes are dominant and localized imputation algorithms such as KNNimpute or LLSimpute perform better than global imputation methods such as BPCA or SVDimpute. On the other hand, if the dataset is more homogenous, a global approach such as BPCA or SVDimpute would better capture the global correlation information in the data. In [27], Jornsten et al. proposes a hybrid approach called LinCmb that captures both global and local correlation information in the data. In LinCmb, the missing values are estimated by a convex combination of the estimates of five different imputation methods: row average, KNNimpute, SVDimpute, BPCA, and GMCimpute. Row average, KNNimpute, and GMCimpute uses local correlation information in their imputation, whereas SVDimpute and BPCA uses global correlation information in their imputation. To obtain the optimal set of weights that combine the five estimates, LinCmb generates fake missing entries at positions where the true values are known and uses the constituent methods to estimate the fake missing entries. The weights are then obtained by performing a least square regression on the estimated fake missing entries. The final weights for LinCmb are obtained by averaging the weights obtained in 30 iterations. LinCmb has been shown to be adaptive to the correlation structure of the

data matrix in that if there are more missing entries in the data matrix, more weights are placed on the global methods.

3.4 Knowledge assisted approach

The common theme for algorithms in this category is the integration of domain knowledge or external information into the imputation process. The use of domain knowledge has the potential to significantly improve the imputation accuracy beyond what is possible with purely data-driven approach, especially for datasets with small number of samples, noisy, or with high missing rate. Algorithms in this category can make use of, for example, knowledge about the biological process in the microarray experiment [28], knowledge about the underlying biomolecular process as annotated in Gene Ontology (GO) [6], knowledge about the regulatory mechanism [39], information about spot quality in the microarray experiment [29], and information from multiple external datasets [63, 23].

Cyclic systems, such as the cell cycle [30] and circadian clock [31] play a key role in many biological processes. Microarray experiments that study these systems are usually carried out by synchronizing a population of cells by first arresting cells at a specific biological life point and then releasing cells from the arrest so that all cells are at the same point when the experiment begins [30, 32]. However, even if cells are perfectly synchronized at the beginning of the experiment, they do not remain synchronized forever [33]. For example, yeast cells seem to remain relatively synchronized for two cycles [30] while wild type human cells lose their synchronization very early [34] or halfway through the first cycle depending on the arrest method. This causes the peak expression value to decrease in amplitude across time. Due to the phenomenon of synchronization loss, the average signal power in successive cycles decreases significantly through time. The biological phenomenon of synchronization loss and the correlation information between genes and between arrays are exploited in a flexible set theoretic framework called Projection onto Convex Sets (POCS) [28] for missing value imputation. In POCS, every known *a priori* property about the original signal is formulated as a corresponding convex set. Given m closed convex sets, POCS successively projects onto these convex sets until reaching a solution at the intersection of all convex sets. POCSimpute captures gene-wise correlation by performing a local least square regression, captures array-wise correlation by PCA imputation, and captures the phenomenon of synchronization loss by restricting the squared power of the expression profiles. POCSimpute is able to adaptively find an optimal solution regardless of whether global or local correlation structure is dominant in the data as the final solution will always be dominated by the smallest (i.e. most reliable) constraint set, while still satisfies the larger (i.e. less reliable) constraint sets.

Functionally related genes tend to express in a modular fashion, with higher degree of concerted reactions to some stimuli [35]. In [6], the a priori information about the functional similarities in term of Gene Ontology (GO) is used for missing value imputation. GO is a well accepted standard for gene function categorization [36, 37] and contains three independent ontologies that describe gene products in terms of their associated biological processes (BP), cellular components (CC) and molecular functions (MF). In GOimpute [6], the semantic similarity is used as the external information on the functional similarity of two genes. GOimpute considers the BP and MF ontologies when calculating the semantic dissimilarity between two genes g_i and g_j. The semantic dissimilarity $d_s(g_i, g_j)$ and the expression level distance $d(g_i, g_j)$ are combined to form a combined distance defined as $c_{ij} = d_s(g_i, g_j)^{\alpha} d(g_i, g_j)$, where the positive weight α controls the relative contribution of the two distance measures. The combined distance is used to select neighborhood genes in KNN and LLS imputation. Experimental results in [6] indicated that GO improves the imputation accuracy when the number of experimental conditions is small or the proportion of annotated genes is large, and at higher rates of missing values. The authors recommended the use of GO information in an imputation if the number of conditions is less than 10 and in particular if the percentage of missing values is sufficiently large (>10%).

Knowledge about the regulatory mechanism can also be used to identify relevant neighbor genes for missing value imputation [39]. It is well known that gene expressions in eukaryotic cells are concertedly regulated by transcription factors and chromatin structure [40]. As histone acetylation could alter chromatin structure and provide binding surfaces for transcription factors, the transcription factor activity is highly regulated by the histone acetylation states in chromatin. It was reported in [41] that genes sharing unique histone acetylation pattern are significantly co-expressed under a wide range of experimental conditions, and are enriched for common biological functions, binding of specific transcription factors, and common DNA motifs. This implies that gene expression activity is strongly regulated by the histone acetylation states [42, 43] and co-expressed genes would have similar acetylation state in their chromatin.

The histone acetylation information aided imputation (HAIimpute) framework proposed in [39] integrates histone acetylation information into KNNimpute and LLSimpute to improve the accuracy of missing value estimation. HAIimpute uses the histone acetylation data of Saccharomyces cerevisiae from [41], where genome wide histone acetylation levels at 11 sites for both intergenic regions (IGRs) and open reading frames (ORFs) were measured. In [41], clusters of genes with similar acetylation patterns across the 11 residues were determined by the *k*-means algorithm. HAIimpute uses the mean expressions of genes from each of the clusters to form the pattern expressions. The missing values in a gene are then obtained by fitting a linear regression model between the gene and the pattern expressions. HAIimpute also perform a secondary imputation using KNNimpute or LLSimpute. The final estimate of the missing value is given by a convex combination of the imputations from linear regression and the secondary imputation. Experimental results in [39] indicated that HAIimpute consistently improve the KNNimpute or LLSimpute and the imputed genes show better correlation with the original complete genes in terms of NRMSE or Pearson correlation.

In [29], information about spot quality in the microarray experiment is taken into account for not only missing value imputation, but also the imputation of unreliable values. The main idea behind the approach is that good quality spots should have more impact on the imputation of other spots, and should themselves be subjected to less imputation than spots with poorer quality. In a typical microarray experiment, a data pre-processing and filtering step is usually taken before any data analysis is performed [44]. In the filtering stage, if a spot has either a small area, low absolute intensity or with average intensity close to background level, with poor SNR, with saturated values in either channel, or a combinations of these, then the spot is usually flagged as unreliable and discarded. This amounts to a binary decision process where a spot is either kept or discarded (missing). One can see that the hard binary decision process is problematic since a spot that is just above the quality threshold will be treated as completely reliable whereas a spot which is just below the threshold will be considered completely unreliable. To take spot quality into account, Johansson and Hakkinen [29] proposed the weighted nearest neighbors imputation (WeNNI). Their method looks at the raw intensity values from the microarray image and compute quality weights *w* for each spot. Then, the imputed value is taken as a linear combination of the original value and the value suggested by its neighbors, i.e. $y_{ik}' = w_{ik} y_{ik} + (1 - w_{ik})\tilde{y}_{ik}$, where $\tilde{y}_{ik}$ takes into account the quality weights of the neighbors and the distance between the neighbors and the original value. Their experimental result on several datasets comparing imputations by row average, KNNimpute, LSimpute_adaptive, and weighted row average shows that WeNNI significantly outperforms the non-weighted algorithms in terms of accuracy as measured by mean squared deviation and robustness due to the inclusion of both spot quality and correlations between genes.

In [45], the idea of using spot quality information in missing value imputation is extended to further consider the case of one-channel depletion. In a two-channel cDNA microarray, missing values could arise due to either a missing reference spot or a missing treatment spot. The case where the spot in one channel is missing while it is measurable in the other channel is called one-channel depletion. For one-channel depletion, the expression ratio of the treatment versus reference cannot be computed and would be flagged as missing. However, it was hypothesized in [45] that if the treatment channel is measurable while the reference channel is un-measurable, it indicates that the treatment sample is up-

regulated with respect to the reference sample and vice-versa. This hypothesis has been observed to be true in five cancer datasets they investigated, where systematic bias (over-estimation in treatment channel depletion and under-estimation in reference channel depletion) was observed in the imputation results using WeNNI. A simple bias correction strategy (WeNNI_BC) involving a constant shift *c* calibrated on non-missing duplicate controls is shown to remove this bias.

In gene expression dataset that has a small number of samples and/or has high missing rate, it is very difficult to reliably identify a set of reference genes from the dataset. One way to overcome this problem is to use multiple external reference datasets in the imputation. The integrative missing value estimation (iMISS) method of [23] integrates information from multiple external reference datasets to improve the imputation. The rationale behind the approach is that if a set of genes frequently exhibit expression similarity to the target gene over multiple external datasets, they constitute a robust neighborhood which tends to show expression co-variations with the target gene. iMISS uses KNNimpute or LLSimpute with average Euclidean distance or rank order statistics for reference genes selection across multiple reference datasets. To assess the appropriateness of the reference datasets, iMISS generates a complete submatrix from the target dataset, then removing 5% of the values randomly from the submatrix to generate 30 evaluation datasets. The imputation algorithms with and without the integrative scheme are then run on the 30 evaluation datasets and their NRMSE are compared by student t-test to assess whether there is any statistical difference between the errors. If the NRMSE of the integrative approach is consistently less than its non-integrative counterpart, the reference datasets are deemed useful for imputation of the target dataset. The experimental result in [23] shows that iMISS consistently performs better than LLSimpute and KNNimpute. In [63], Jornsten et al. propose using an external database matrix generated from publicly available microarray data as meta-data for imputation. In their method (denoted here as metaMISS), the database matrix of a species is obtained from all available microarrays of the species downloaded from the Stanford Microarray Database (SMD) [64]. Given a gene expression dataset *A*, the missing values in a column in *A* are imputed by first finding a set of 40 most similar columns from the database matrix or from *A* as measured by the absolute Pearson correlation. GMCimpute is then used to impute the missing values in the column from this set of 40 most similar columns. Their results indicated that better imputation accuracy in terms of NRMSE can be obtained with this approach.

4. Validation of Imputation Results

Validation of imputation results is an important step in assessing the performance of imputation algorithms. In general, validation is done by computing certain performance indices between the imputed and the (known) original values. We called this kind of validation internal validation, i.e. the validation uses only information in the dataset. Examples of performance indices for internal validation include (i) fidelity to true expression values, (ii) preservation of internal structures in the dataset, (iii) preservation of significant genes in the dataset, and (iv) preservation of discriminative/predictive power for classification. Internal validation is by far the most commonly used validation method for missing value imputation. Nevertheless, in view of the fact that missing value imputation is to facilitate downstream analysis, validation can also be done by assessing the effect of imputation on subsequent biological analysis. We called this external validation, i.e. the validation uses external knowledge. External validation is less straightforward to perform and usually suffer from lack of relevant information (e.g. incomplete GO annotation or functional annotation), but it is more relevant to the final goal of missing value imputation which is to facilitate downstream biological analysis. Hence, external validation should be performed whenever practicable. Table 2 lists the different validation methods that have been used by different researchers.

<< Table 2 here >>

4.1 Internal validation

(i) Fidelity to true expression values

The most common method to assess imputation accuracy is to compute the normalized root mean square error (NRMSE) or variants of it (e.g, RMSE, where the denominator in (1) is replaced by *mn*) between the imputed values and the true values [9-12, 14-20, 23, 26-28, 39, 46, 48]. The NRMSE is defined as

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{m}\sum_{k=1}^{n}(g_{ik} - \tilde{g}_{ik})^2}{\sum_{i=1}^{m}\sum_{k=1}^{n}(g_{ik})^2}} \qquad (1)$$

where g_{ik} denotes the value of the *k-th* experiment for gene g_i, and g and $\tilde{g}$ denote the true value and the imputed value respectively. To overcome the non-availability of ground truth, assessment of imputation algorithm is usually done on the complete data matrix derived from the original data matrix. In deriving the complete data matrix, genes with missing values are excluded. Then missing values are randomly generated on the complete data matrix either assuming a uniform distribution of missing entries or according to some distribution models of missing entries estimated from the actual dataset. Imputation algorithms with lower NRMSE are more accurate at imputing the missing values. Note that since NRMSE effectively calculates the second moment of the residual, i.e. it assumes that the residual has a normal distribution, this performance measure would not be appropriate for algorithms that deviate from this normality assumption.

When replicates are available, i.e. when microarray experiments are performed in replicates, the imputed values can be compared with the replicates (which serve as ground truth) to assess the imputation accuracy. In [29], the accuracy of imputation is assessed using the mean squared deviation (MSD)

$$MSD = \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n}(\tilde{x}_{ik} - \gamma_{ik})^2 \qquad (2)$$

where $\tilde{x}$ is the imputed data, γ is the replicate data, and the summation is run over all expression values in the replicate dataset except where they are marked as invalid in the data pre-processing step. One can see that except for the choice of ground truth, the MSD is the same as the squared of the RMSE.

In [16], the data matrix is log-transformed before missing value imputation is performed. The resulting RMSE of the imputation of the log-transformed data is denoted as log-transformed RMSE (LRMSE). As log-transformation improves the normality of data distribution and is scale-invariance to power law distribution, it could allow better comparison of imputation accuracy across different datasets. However, as negative expression values are removed prior to log-transformation in [16], less data are used for imputation and this could affect imputation performance. The log-transformation of the data could also favors imputation algorithms that have a normality assumption in the data distribution or algorithms that employ the Euclidean distance-based metric. Hence, the performance of imputation algorithms and their relative ranking could potentially be affected by this data pre-processing step. Note that values in a gene expression data matrix are usually given by the log-transformed of the raw gene expression ratios, the effect of an additional log-transformation of the dataset prior to imputation warrants further investigation.

(ii) Preservation of internal structures

Another way to assess the performance of an imputation algorithm is by measuring how well the clustering of the complete dataset was preserved when clustering the imputed dataset [7, 47, 48]. This type of assessment looks at how well the internal structure of the dataset (as measured by a clustering algorithm) is preserved by the imputation method. It has been shown that even if two imputation methods differ markedly in their NRMSE, their differences become negligible when they are evaluated in terms of how well they can reproduce the original gene clusters [47]. In [47], the

agreement between the partitioning of the imputed dataset and the complete dataset by K-mean clustering is measured by the average distance between partitions (ADBP). The method works by first matching optimally each of the K clusters in the partitioning of the imputed dataset with its best pair in the K partitioning of the complete dataset before measuring the dissimilarity between the two partitions. Let U and V be two different partitions of a dataset. The normalized Hamming distance between a cluster pair $\{u_j \in U, v_j \in V\}$ is given by

$$d_h(u_j, v_j) = \frac{1}{n(u_j) + n(v_j)} \left(\sum_{g \in u_j} I\{g \notin v_j\} + \sum_{g \in v_j} I\{g \notin u_j\} \right) \qquad (3)$$

where $n(u_j)$ and $n(v_j)$ are the cardinalities of the clusters u_j and v_j, respectively, and $I\{g \notin u_j\}$ equals one if gene g does not belong to the cluster u_j, and zero otherwise [49]. Every possible combination of $u_i \in U$, $v_j \in V$ pair would generate a $K \times K$ matrix of $d_h(u_i, v_j)$. Then the optimum K cluster pairs can be obtained from the $K \times K$ matrix by solving an assignment problem using the Hungarian algorithm [50, 51]. Finally, given the optimum k cluster pairs $\{ u_j \in U, v_j \in V\}$, the ADBP error is calculated as

$$D(U, V) = \frac{1}{K} \sum_{j=1}^{K} d_h(u_j, v_j) \qquad (4)$$

Imputation algorithm that produces a lower ADBP error could preserve the structure within the dataset better than algorithm that has a higher ADBP error.

In [7], the performances of no imputation, KNNimpute, and replacement by zero are assessed in terms of how well the imputation preserves the hierarchical clustering of the complete dataset. The hierarchical clustering results from the complete dataset and from the imputed dataset are compared using the conserved pairs proportion (CPP) index which measures the percentage of conserved genes between the clusters of the reference set and the imputed set. To compute the CPP between the reference clusters and the imputed clusters, one first builds a contingency table with rows denoting the reference clusters and columns denoting the imputed clusters. Then the entries in the contingency table are filled in by counting the number of genes that are common to the clusters indexed by the row and column. Finally the CPP is computed by summing the largest entry in every row and divided by the total number of genes. It was observed in [7] that missing values have significant effects on the stability of the gene clusters. Even a 1% missing rate is enough to cause genes to be relocated to different clusters in the complete-linkage hierarchical clustering result. With the missing values imputed by KNNimpute, the bias in the clustering caused by missing values can be significantly reduced. The idea of comparing the performance of imputation algorithms based on how well they preserve the cluster structure using CPP is also taken up in a recent comparative study of many existing imputation algorithms in [48].

(iii) Preservation of significant genes

Ideally, a good imputation algorithm should preserve interesting genes in the dataset, while minimizes the chance of artificially inflating genes that were originally non-differentially expressed [14, 27, 52]. In [52], the change in the list of differentially expressed genes detected due to imputation is used as a performance measure for imputation algorithms. The performance is measured by checking how many differentially expressed genes are lost or added to the original list of differentially expressed genes that is obtained from the full dataset due to imputation errors. To detect differentially expressed genes, [52] uses the ANOVA method of [53] which is based on a linear model of gene expression, as well as an approach based on hypothesis testing directly on the matrix of log ratios as done by Significance Analysis of Microarrays (SAM) [54]. The SAM test is also used in [14] to determine the percentage of lost differentially expressed genes after imputation. In [27], ANOVA and standard or regularized t-test with correction for multiple testing are used to test for significant differential expressions in the complete dataset to obtain the gold standard. Then, standard or regularized t-test is

applied to the imputed dataset and the list of genes sorted by the *P*-value is compared to the gold standard to compute the false positive rate at a false negative rate of 0% and 5%.

(iv) Preservation of discriminative/predictive power

The performance of imputation algorithms can also be assessed by how the imputation affects the classification accuracy of the imputed dataset for different disease types [22, 46, 55]. The study in [55] looks at the effects of missing values and their imputation on classification performance. Three imputation algorithms (KNNimpute, LLSimpute, and BPCA) are applied to five different cancer datasets and the classification accuracy is assessed by three different classifiers (SVM, KNN, classification and regression tree (CART)). It was shown that except for replacement by zeros, other imputation algorithms have little difference in affecting classification performances of the SVM or KNN classifiers. In [46], the impact of missing value imputation on classification accuracy is studied by using synthetically generated datasets. Moreover, rather than random generation of missing values, the missing values are generated by identifying and discarding the low quality entries in the synthetic dataset. Six imputation algorithms, namely, row average, KNNimpute, LLSimpute (both L_2 and Pearson correlation based), LSimpute, and BPCA, are studied with respect to how well the imputed dataset can preserve the discriminative power in the original dataset as measured by three different classifiers (LDA, KNN, and linear SVM). Their results suggested that it is beneficial to apply missing value imputation when the noise level is high, variance is small, or gene-cluster correlation is strong, under small to moderate missing rates. In [22], genes from imputed dataset are ranked by their between sum of squares to within sum of squares (BSS/WSS) ratio, which measures the ability for class prediction. The *p* most significant genes are selected and compared with the *p* most significant genes selected from the complete dataset in terms of class prediction capability using the true positive rate score.

Other performance measure that has been used to validate or compare performance of imputation algorithms includes Pearson correlation between imputed and original data [14, 39]. Pearson correlation measures the strength of linear association between two variables. In [14], the capability of the imputation method to retain the structure of each experiment (column) is determined by the Pearson correlation coefficient R^2 between the estimated values and the true values for each column of the data matrix. In [39], the Pearson correlation between the imputed genes and the genes in the complete dataset is calculated to assess the accuracy of the imputation.

4.2 External validation

External validation assesses the merit of an imputation algorithm by examining how it affects downstream biological analysis with the aid of external information related to functional annotation or pathway information. Different criteria, such as functional enrichment of gene clusters [47], and improvement in biological interpretation of the data [22], have been proposed for validation.

In many microarray experiments, the goal is to detect sets of genes that share similar functional roles. Once clustering has been done, a biologist typically looks for GO terms that are significantly enriched among the genes of each cluster [47, 56-58]. The GO terms can then be used to characterize the functional roles of the genes under study. Hence, one way of assessing the performance of an imputation algorithm is to check for significant enrichment of GO terms in gene clusters. In [47], the enrichment *p*-value for each GO terms *t* and for each cluster is calculated by

$$p = \sum_{i=K}^{\min(b,T)} \frac{\binom{T}{i}\binom{B-T}{b-i}}{\binom{B}{b}} \tag{5}$$

where b is the number of genes in the cluster, K is the number of genes in the cluster annotated with GO term t, B is the number of genes in the dataset, and T is the number of genes in the dataset annotated with GO term t. The p-value in (20) expresses the probability of randomly finding the same or higher number of genes annotated with the particular GO term in the cluster from the dataset. In [47], between one to 20 most enriched GO terms for each cluster of the reference partition and the imputed partition were obtained and these clusters of GO terms were compared using the ADBP measure to quantify the proportion of GO terms that were present in the complete dataset and were preserved in the imputed dataset. One interesting finding is that even when there are marked differences in the NRMSE among different imputation algorithms, these differences become negligible when the methods are evaluated in terms of how well they can reproduce the original gene clusters or their biological interpretations with GO.

In [22], the different imputation algorithms are compared based on whether biologically important genes still remain significant after imputation. This comparison is particularly relevant to studies where the aim is to detect marker genes for disease characterization. A good imputation algorithm should preserve the significance of the marker genes after missing value imputation. To assess the performance of the imputation algorithms, the set of significant genes selected by using the BSS/WSS ratio are examined for any known biologically important genes. Their results on the breast cancer dataset show that important genes like KIAA1025 (co-regulated with estrogen receptor for both in vivo and clinical data, and is expressed in more than 66% of human breast tumors [59]) and PKP2 (found in breast carcinoma cell lines [60] and could serve as a marker for the identification and characterization of carcinomas derived either from or corresponding to simple or complex epithelia [61]) could be missed by popular imputation algorithms like LLSimpute, BPCA, and KNNimpute at even low (1%) to moderate (5%) missing rate.

5. Factors Affecting Imputation Performance

It has been shown in several studies that the performance of imputation algorithms is significantly affected by factors such as the correlation structure in the data, the missing-data mechanism, the distribution of missing entries in the data, and the percentage of missing values in the data [16, 46, 47, 52]. Hence, there is no one imputation algorithm that performs the best in every situations and choosing the right algorithm may significantly boost the accuracy of the imputation results.

In [16], the performances of eight different imputation algorithms are compared on several gene expression datasets involving time series, multiple exposure (i.e. non time series dataset such as those from cancer study), and mixture of both types. To measure the underlying correlation structure in the dataset, an entropy measure of the dataset D, given by

$$e(D) = -\frac{\sum_{i=1}^{k} p_i \log p_i}{\log k} \tag{6}$$

where $p_i = \sqrt{\lambda_i} / \sum_{i=1}^{k} \sqrt{\lambda_i}$, and λ_i , i =1, …, k, are the eigenvalues of the covariance matrix of the data, is used to measure the dispersion in the eigenvalues. Low entropy indicates that the data are strongly correlated. In contrast, high entropy indicates a complex data exhibiting strong local substructure. As expected, global methods, such as SVDimpute and BPCA, perform better on microarray datasets with low entropy and local methods, such as LLSimpute and KNNimpute, perform better on high entropy datasets.

The missing-data mechanism has an important bearing on the performance of missing value imputation. In [65], three mechanisms of missingness are defined: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). An observation is said to be MCAR if the missingness is independent of the observed and unobserved data. An observation is said to be MAR if the missingness is independent of the unobserved data, although it may be

dependent on the observed data. An observation that is neither MCAR nor MAR is called missing not at random (MNAR). MNAR occurs if missingness depends not only on the observed data but also on the unobserved (missing) values. In most studies, missing values in microarray datasets are assumed to be missing at random and most imputation algorithms either explicitly or implicitly make this assumption. However, this is not a realistic assumption since missing values tend to arise in a systematic manner in practice. In a typical microarray experiment, if a spot has low intensity in any one of the channels, i.e. its foreground intensity is close to its background intensity in the channel, it is usually flagged as unreliable and discarded. In this case, the missingness is actually dependent on the signal intensity, and hence is MNAR. Statistical inferences on data with MNAR generally lead to bias estimates [65]. It has been shown experimentally that missing value imputation algorithms generally tend to give poorer result in terms of NRMSE when missing value is MNAR [52]. It was observed that for the same percentage of missing values, the percentage lost in differentially expressed genes is higher for MNAR than for MAR, and imputing at 5% missing rate when MNAR has the same effect as imputing at 10%-30% missing rate when MAR [52].

The distribution of missing entries in the data matrix, i.e., the missing-data pattern, can also affect imputation performance and should be considered in data analysis or algorithm design [65]. Since each column in a microarray data matrix comes from one microarray experiment, variation in experimental conditions across the columns could result in a non-random distribution of missing entries in the data matrix. In [47], the effect on the imputation accuracy when missing values are randomly distributed in the data matrix and when the distribution of the missing entries follows that of the actual distribution of missing entries are compared. Preliminary imputation experiment on a time series yeast cell cycle dataset indicates that the NRMSE tends to be higher for the case of randomly distributed missing values, especially at low missing value rate. In [39], it was observed empirically in several datasets that missing values in genes occur successively, i.e. in burst. The performance of imputation algorithms were observed to be somewhat lower, especially at low missing rate, for burst model of missing entries compare to missing at random.

The effect of missing rate on imputation has been investigated in [46]. In [46], the performance of six popular imputation algorithms was assessed at missing rate of 1, 5, 10, 15, 20, and 25%. It was observed that there is improvement on the performance, as measured by the NRMSE, of the various imputation methods studied as the missing rate initially increases and then the performance start to deteriorate beyond a certain missing rate. This is in contrast to many other studies which show that the NRMSE generally increases monotonically with increasing missing rate. The authors suggested that this could be a consequence of the fact that in their test datasets, the missing values are generated non-randomly based on a spot quality threshold.

6. Discussion and Future Directions

In spite of the many recent advances, better imputation algorithms that can adapt to the characteristics of the data are still needed. Adaptive method that could capture both global and local correlation information in the dataset would be useful in many situations. One attempt in this direction is the hybrid approach LinCmb [27] we discussed in Section 3.3. The adaptive weighting of the estimates from a set of global and local imputation algorithms allows LinCmb to combine the various estimates and hence adapts to the correlation structure in the data. Nevertheless, questions such as deciding on the set of imputation algorithms to be included in the linear combination and how to efficiently compute the weights still need further research. It is also unclear whether the simple scheme of linearly combining the estimates from multiple imputation algorithms is the best strategy to adopt or some other more sophisticated strategy is needed. One possible combination strategy is to combine multiple estimates based on the variance in each of the estimates. This requires the estimation of the variance of each imputation, which could be done by a procedure called multiple imputation [65].

Recently, we have investigated an integrated framework of performing missing value imputation and bicluster analysis. The integrated framework is based on the fact that both missing value imputation and bicluster analysis rely on exploiting correlation information within the data and should be

considered together instead of independently. The integrated framework iterates between missing value imputation and bicluster analysis, using the result of one process to update and improve the result of the other process. The integrated framework means that missing value imputation can benefit from information gained from bicluster analysis and vice versa. For example, if the biclusters detected are of a multiplicative model [56, 57], the multiplicative model can be used as a prior for missing value imputation, thus improving the imputation accuracy. The better imputation would in turn leads to better bicluster analysis in subsequent iteration. Our preliminary results have shown that this integrated framework is promising, although there is still much research to be done here.

Several recent studies [16, 48] have attempted to evaluate the relative performance of existing imputation algorithms but there is still no clear consensus about the best algorithm for every datasets. This is partly due to the fact that there are many factors that can affect the performance of an imputation algorithm (see discussion in Section 5) and there is probably no algorithm that performs the best in all datasets. Extensive large scale systematic comparative studies, with carefully designed datasets of different nature and properties as well as experimental design that takes into account the various factors that could impact performance, are still very much needed. The very recent work in [48] which runs more than six millions simulations over five datasets comparing the performance of 12 imputation algorithms in terms of NRMSE is a step toward this goal.

The performance assessment of existing imputation algorithms with respect to the different datasets as discussed above can be viewed as data-driven assessment. It should be noted that the choice of imputation algorithm also depends very much on the applications at hand. Hence, the performance of imputation algorithms should also be assessed from an application-driven viewpoint. As one reviewer pointed out, simple imputation technique such as KNNimpute might suffice for many applications even though it has lower NRMSE than most advanced imputation techniques. Nevertheless, imputation algorithm with high accuracy is desirable and even necessary for certain applications. For example, we found that if the estimated missing values of the alpha factor yeast dataset using the spectral estimation method of [25] are perturbed randomly by addition a zero mean Gaussian noise of increasing magnitude to the estimated values, the number of detected cyclic genes (within the top 400 ranking cyclic genes) that overlaps with the benchmark set of known cyclic genes B1 tends to decrease in general. A 5%, 10%, and 20% random perturbation of the imputed missing values causes the overlap to decrease by 0.3%, 2.4%, and 10%, respectively. While the decrease is not significant at 5% level (only ~1 known cyclic gene is missed) and classical imputation techniques would suffice in this case, highly inaccurate missing value estimation can adversely affect the detection of cyclic genes (i.e. almost 10 cyclic genes are missed at the 10% level). To date, we are not aware of any comprehensive, large scale comparative study of imputation performance of existing algorithms from an application-driven point of view. Clearly, comparing the performance of existing imputation algorithms in terms of the end applications would be beneficial to the end users.

As more and more experimental data from different domains become available, new imputation algorithms that can handle mixed domain datasets with missing continuous and categorical data are needed. Mixed domain datasets could, for example, arise from studies that simultaneously considering microarray gene expression data (i.e. continuous data) and classes of known phenotypic variables such as clinical chemistry evaluations and histopathologic observations (i.e. categorical data) [66]. Missing value imputation in mixed domain datasets is a very challenging problem and may require the use of highly sophisticated statistical techniques as discussed in [65].

7. Conclusion

High throughput gene expression profiling techniques such as cDNA microarray technology usually suffer from missing value problem due to various experimental reasons. As many downstream analysis methods require a complete dataset, missing value imputation is an important pre-processing step in microarray data analysis. Troyanskaya et al. [9] were the first to bring this important problem to the attention of the bioinformatics research community. Since then, many missing value imputation

algorithms have been proposed [6, 9-15, 17-23, 26-29, 39, 45, 52, 63]. In this paper, we present a comprehensive review of many of these algorithms, categorizing them in terms of how they utilize information from within the data or information from domain knowledge or external sources in the imputation. In our discussion, we group different imputation algorithms into local approach, global approach, hybrid approach, or knowledge assisted approaches. We provide a succinct description of the underlying methodology of each algorithm regarding the rationale, the technical framework, and the assumptions behind them. In this way, the connection between different imputation methods becomes more apparent to researchers. Validation of imputation result is an important step in assessing the performance of any imputation algorithm. To date, there is a lack of any systematic discussion of this important issue. In this paper, we summarize the various validation methods used by different researchers and categorized them into either internal or external validation, based on whether internal information from within the dataset or external biological knowledge is used. We also discuss some factors that could affect imputation performance. These factors should be considered when choosing an imputation method for a particular dataset. Finally, we conclude the survey by pointing out some possible future research directions. It is hoped that this comprehensive review would give the readers a better understanding of the current development in this field and inspire them to come up with the next generation of imputation algorithms.

Funding

Hong Kong Research Grant Council (Projects CityU123408 and CityU123809).

References

1. Hoheisel JD. Microarray technology: beyond transcript profiling and genotype analysis. *Nat. Rev. Genet.* 2006; **7**:200-210.
2. Armstrong SA, Staunton JE, Silverman LB, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.* 2002; **30**: 41-47.
3. Muro S, Takemasa I, Oba S, et al. Identification of expressed genes linked to malignancy of human colorectal carcinoma by parameteric clustering of quantitative expression data. *Genome Biol.* 2003; **4**: R21.
4. Kim S, Dougherty ER, Chen Y, et al. Multivariate measurement of gene expression relationships. *Genomics* 2000; 67, 201-209.
5. Duggan DJ, Bittner M, Chen Y, et al. Expression profiling using cDNA microarrays. *Nat. Genet.* 1999; 21: 10-14.
6. Tuikkala J., Elo L., Nevalainen O., et al. Improving missing value estimation in microarray data with gene ontology. *Bioinformatics* 2006; 22(5): 566–572.
7. 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.
8. Yang YH, Buckley MJ, Dudoit S, et al. Comparison of methods for image Analysis in cDNA microarray data. Technical Report 584, Dept. of Stat., UC Berkeley, 2000.
9. Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001; 17(6):520-525.
10. Oba S, Sato MA, Takemasa I, et al. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* 2003; 19(16):2088-2096.
11. 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.
12. Kim H, Golub GH, Park H. Missing Value Estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics* 2005; 21(2):187-198.
13. Kim KY, Kim BJ, Yi GS. Reuse of imputed data in microarray analysis increases imputation efficiency. BMC Bioinformatics 2004; 5:160
14. Bras LP, Menezes JC. Improving cluster-based missing value estimation of DNA microarray data. *Biomolecular Engineering* 2007; 24: 273-282.
15. Ouyang M, Welsh WJ, Georgopoulos P. Gaussian mixture clustering and imputation of microarray data. *Bioinformatics* 2004; 20(6):917-923.
16. Brock GN, Shaffer JR, Blakesley RE, et al. Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. *BMC Bioinformatics* 2008; 9:12.

17. Zhang X, Song X, Wang H, et al. Sequential local least squares imputation estimating missing value of microarray data. *Computers in Biology and Medicine* 2008; 38:1112-1120.
18. Cai Z, Heydari M, Lin G. Iterated local least squares microarray missing value imputation. *J Bioinform Comput Biol.* 2006; 4(5):935-57.
19. Yoon D, Lee EK, Park T. Robust imputation method for missing values in microarray data. BMC *Bioinformatics* 2007; 8 (Suppl 2):S6.
20. Zhou X, Wang X, Dougherty ER. Missing value estimation using linear and non-linear regression with Bayesian gene selection. *Bioinformatics* 2003; 19(17): 2302-2307.
21. Sehgal MS, Gondal I, Dooley LS. Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. *Bioinformatics* 2005; 21(10):2417-2423.
22. Sehgal MS, Gondal I, Dooley LS, et al. Ameliorative missing value imputation for robust biological knowledge inference. *Journal of Biomedical Informatics* 2008; 41: 499-514.
23. Hu J, Li H, Waterman MS, et al. Integrative missing value estimation for microarray data. BMC *Bioinformatics* 2006; 7:449.
24. Bar-Joseph Z, Gerber GK, Gifford DK, et al. Continuous representation of time series gene expression data. *Journal of computational biology* 2001; 10: 341-356.
25. Liew AWC, Xian J, Wu S, et al. Spectral estimation in unevenly sampled space of periodically expressed microarray time series data. *BMC Bioinformatics* 2007; 8:137.
26. Choong MK, Charbit M, Yan H. Autoregressive model based missing value estimation for DNA microarray time series data. *IEEE Trans. Information Technology in Biomedicine* 2009; 13(1):131-137.
27. Jornsten R, Wang HY, Welsh WJ, et al. DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics* 2005; 21(22):4155-4161.
28. 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):1608-1619.
29. Johansson P, Hakkinen J. Improving missing value imputation of microarray data by using spot quality weights. *BMC Bioinformatics* 2006; 7(1):306.
30. Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 1998; 9:3273-3297.
31. Panda S, Antoch MP, Miller BH, et al. Coordinated transcription of key pathways in the mouse by the circadian clock. Cell 2002; 109:307–320.
32. Gasch AP, Spellman PT, Kao CM, et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 2000; 11:4241-4257.
33. Bar-Joseph Z, Farkash S, Gifford DK, et al. Deconvolving cell cycle expression data with complementary information. *Bioinformatics* 2004; 20:i23–i30.
34. Shedden K, Cooper S. Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proc. Natl Acad. Sci. USA* 2002; 99: 4379–4384.
35. Hartwell LH, Hopfield JJ, Leibler S, et al. From molecular to modular cell biology. *Nature* 1999; 402:C47–52.
36. Consortium TGO. Gene Ontology Tool for the Unification of Biology. *Nat. Genet*. 2000; 25:25–29.
37. Draghici S, Khatri P, Martins RP, et al. Global functional profiling of gene expression. *Genomics* 2003; 81(2):98–104.
38. Carey VJ. Ontology concepts and tools for statistical genomics. *Journal of Multivariate Analysis* 2004; 90(1):213–228.
39. Xiang Q, Dai X, Deng Y, et al. Missing value imputation for microarray gene expression data using histone acetylation information. *BMC Bioinformatics* 2008; 9:252.
40. Yuan GC, Ma P, Zhong W, et al. Statistical assessment of the global regulatory role of histone acetylation in Saccharomyces cerevisiae. *Genome Biology* 2006; 7:R70.
41. Kurdistani SK, Tavazoie S, Grunstein M. Mapping global histone acetylation patterns to gene expression. *Cell* 2004; 117(6):721-733.
42. Verdone L, Caserta M, Di Mauro E. Role of histone acetylation in the control of gene expression. *Biochem Cell Biol* 2005; 83(3):344-353
43. Guo X, Tatsuoka K, Liu RX. Histone acetylation and transcriptional regulation in the genome of Saccharomyces cerevisiae. *Bioinformatics* 2006; 22(4):392-399.
44. Liew AWC, Yan H, Yang M. Pattern recognition techniques for the emerging field of bioinformatics: a review. *Pattern Recognition* 2005; 38:2055-2073.
45. Ritz C, Eden P. Accounting for one-channel depletion improves missing value imputation in 2-dye microarray data. *BMC Genomics* 2008; 9:25
46. Sun Y, Braga-Neto U, Dougherty ER. Impact of Missing Value Imputation on Classification for DNA Microarray Gene Expression Data—A Model-Based Study. *EURASIP Journal on Bioinformatics and Systems Biology* 2009:504069.

47. Tuikkala J, Elo LL, Nevalainen OS, et al. Missing value imputation improves clustering and interpretation of gene expression microarray data. *BMC Bioinformatics* 2008; 9:202.
48. Celton M, Malpertuy A, Lelandais G, et al. Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. *BMC Genomics* 2010; 11:15.
49. Lange T, Roth V, Braun ML, et al. Stability-based validation of clustering solutions. *Neural Computation* 2004; 16:1299-1323.
50. Kuhn HW. The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly* 1955; 2:83–97.
51. Burkard RE, Dell'Amico M, Martello S. Assignment Problems. SIAM, Philadelphia (PA), 2009.
52. Scheel I, Aldrin M, Glad IK, et al. The influence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics* 2005; 21(23):4272-4279.
53. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. *J. Comput. Biol.* 2000; 7(6):819–837.
54. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* 2001; 98(9):5116–5121.
55. Wang D, Lv Y, Guo Z, et al. Effects of replacing the unreliable cDNA microarray measurements on the disease classification based on gene expression profiles and functional modules. *Bioinformatics* 2006; 22(23):2883–2889.
56. Cheng KO, Law NF, Siu WC, et al. Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization. *BMC Bioinformatics* 2008; 9:210
57. Gan X, Liew AWC, Yan H. Discovering biclusters in gene expression data based on high-dimensional linear geometries. *BMC Bioinformatics* 2008; 9:209.
58. Zhao H, Liew AWC, Xie X, et al. A new geometric biclustering algorithm based on the Hough transform for analysis of large-scale microarray data. *Journal of Theoretical Biology* 2008; 251:264–274
59. Harvell DME, Richer JK, Allred DC, et al. Estradiol regulates different genes in human breast tumor xenografts compared with the identical cells in culture. *Endocrinology* 2006; 147:700–13.
60. Mertens C, Kuhn C, Franke W. Plakophilins 2a and 2b: constitutive proteins of dual location in the karyoplasm and the desmosomal plaque. *J Cell Biol* 1996; 135:1009–25.
61. Mertens C, Kuhn C, Moll R, et al. Desmosomal plakophilin 2 as a differentiation marker in normal and malignant tissues. *Differentiation* 1999; 64:277–90.
62. Aittokallio T. Dealing with missing values in large-scale studies: microarray data imputation and beyond. Briefings in Bioinformatics 2010; 11:253-264.
63. Jornsten R, Ouyang M, Wang HY. A meta-data based method for DNA microarray imputation. *BMC Bioinformatics* 2007; 8:109.
64. Sherlock G, Hernandez-Boussard T, Kasarskis A, et al. The Stanford Microarray Database. *Nucleic Acids Res* 2001; 29:152-5.
65. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd Ed., Wiley Interscience, John Wiley & Sons, Inc, New Jersey, 2002.
66. Baskin CR, García-Sastre A, Tumpey TM, et al. Integration of clinical data, pathology, and cDNA microarrays in influenza virus-infected pigtailed macaques (*Macaca nemestrina*). J Virol. 2004; 78(19): 10420–10432.

Key Points:

- Global correlation information, local similarity information, and domain knowledge have been successfully exploited in many existing missing value imputation algorithms.
- The performance of an imputation algorithm can be assessed using internal data-driven validation and external domain/application-specific validation.
- There is no one optimal imputation algorithm for all type of data. Knowledge of the underlying principle of the imputation algorithm and the characteristic of the data is needed for a good imputation performance.

About the authors:

Alan Wee-Chung Liew is an Associate Professor at the School of Information and Communication Technology, Griffith University, Australia. His research in bioinformatics focuses on developing novel computational techniques for gene expression data and DNA sequence analysis. Dr. Liew is a senior member of IEEE.

Ngai-Fong Law is an Associate Professor at the Centre for Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong. Her research focuses on applying advanced signal processing techniques to problems in bioinformatics. Dr. Law is a senior member of IEEE.

Hong Yan is a Professor at the Department of Electronic Engineering, City University of Hong Kong, Hong Kong. His research in bioinformatics focuses on gene expression data and DNA sequence analysis using advanced signal processing and pattern recognition techniques. Prof Yan is a Fellow of IEEE and Fellow of IAPR.