BRIEFINGS IN BIOINFORMATICS. VOL 11. NO 2. 253–264
Advance Access published on 4 December 2009

doi:10.1093/bib/bbp059

Dealing with missing values in large-scale studies: microarray data imputation and beyond

Tero Aittokallio

Submitted: 1st October 2009; Received (in revised form): 1st November 2009

Abstract

High-throughput biotechnologies, such as gene expression microarrays or mass-spectrometry-based proteomic assays, suffer from frequent missing values due to various experimental reasons. Since the missing data points can hinder downstream analyses, there exists a wide variety of ways in which to deal with missing values in large-scale data sets. Nowadays, it has become routine to estimate (or impute) the missing values prior to the actual data analysis. After nearly a decade since the publication of the first missing value imputation methods for gene expression microarray data, new imputation approaches are still being developed at an increasing rate. However, what is lagging behind is a systematic and objective evaluation of the strengths and weaknesses of the different approaches when faced with different types of data sets and experimental questions. In this review, the present strategies for missing value imputation and the measures for evaluating their performance are described. The imputation methods are first reviewed in the context of gene expression microarray data, since most of the methods have been developed for estimating gene expression levels; then, we turn to other large-scale data sets that also suffer from the problems posed by missing values, together with pointers to possible imputation approaches in these settings. Along with a description of the basic principles behind the different imputation approaches, the review tries to provide practical guidance for the users of high-throughput technologies on how to choose the imputation tool for their data and questions, and some additional research directions for the developers of imputation methodologies.

Keywords: *missing value imputation; gene expression microarrays; mass-spectrometry proteomics; statistical modelling; biomarker discovery; disease classification*

INTRODUCTION

The problems posed by missing observations are well established in statistical data analysis literature [1]. The standard statistical methods have been developed to analyse complete data matrices, in which the rows represent cases and the columns are variables measured for each case; however, in many applications, there are entries of the data matrix which are not observed. For instance, variables for some of the cases cannot be measured due to technical problems unrelated to the experimental question (this is called 'missing completely at random' because the reason for missingness is totally random), or the measurements are not reliable or obtainable for some particular cases (this is called, somewhat confusingly, 'missing at random' because the missing data can be considered a random sample conditional on the other observed characteristics of the cases that determined their missingness). When the number of missing values in a data set is large, it is not reasonable to simply discard such observations or remove the corresponding cases, since this will lose valuable information and can lead to selection bias; instead, the missing values need to be replaced or predicted as accurately as possible before the actual data analysis. Filling missing values with zeros or

Corresponding author. Tero Aittokallio, Biomathematics Research Group, Department of Mathematics, FI-20014 University of Turku, Finland. Tel: +358-2-333-6030; Fax: +358-2-333-6595; E-mail: tero.aittokallio@utu.fi

Tero Aittokallio is an Academy Research Fellow in the Biomathematics Research Group at the Department of Mathematics, University of Turku, Finland. He is developing experimental–computational approaches for problems in systems biology, with special focus on large-scale data mining and modelling in medical research.

with average values over the cases are far from optimal solutions, and generally lead to serious biases, as they do not take into consideration the correlation structure in the data. Therefore, a number of more sophisticated statistical models and procedures that efficiently use the information captured in the non-missing part of the data set to estimate (or 'impute') missing data values have been developed and successfully applied in several data-rich application areas, especially in medical research [2].

In the context of microarray technology, it soon became evident after the first applications to larger-scale gene expression profiling studies [3–5], that it is essential to somehow deal with the frequent missing data points that are inherent to these high-throughput assays. One practical problem is that many standard methods for gene expression data analysis require a complete data matrix as an input. In a sense, missing value imputation can be considered as a cost-effective alternative to repeating all those experiments with missing data points. While the earliest imputation methods were rather straightforward applications of the standard statistical imputation approaches to the microarray data sets, recently more application-specific modifications have been introduced that take advantage of the particular properties of the data being imputed and possibly also exploit other information sources relevant to the imputation task. After nearly a decade since the first applications of imputation methods to gene expression microarray data, it seems that the rate of publications on missing data imputation is not slowing down; on the contrary, the methodologists are developing new and improved methods at an increasing rate (Fig. 1). This is not surprising as the problem provides a rich source of interesting questions for the computational groups entering the field of bioinformatics and computational biology. For the more biologically oriented researchers, the prediction of expression levels from multiple information sources, such as functional annotations, histone acetylation or transcription-factor-binding information, also offers a systematic computational means to investigate the interrelationships among the various data sources, and their contribution to gene expression patterns and dynamics.

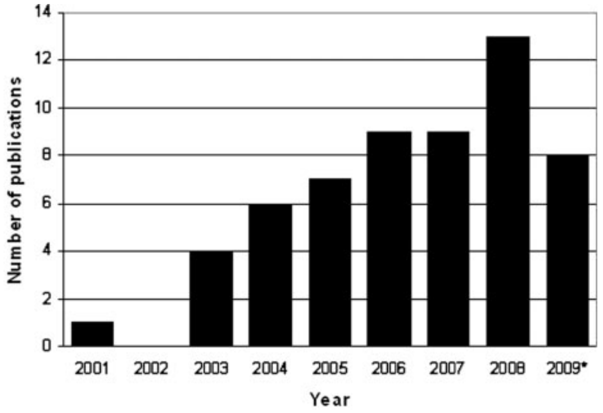

Figure 1: The number of publications found in PubMed on the topic of missing value imputation in microarray data. Asterisk indicates number of publications as of Ist of October, 2009.

Although a number of the imputation algorithms have been made available to the experimental practitioners, and some of them are also easily accessible through software tools, only limited efforts have been devoted to providing the users of the high-throughput platforms with practical guidance on how to choose between the different methods, or how to optimize their performance for a given data set in real applications. From the user's point of view, rather than introducing incremental improvements to the standard imputation approaches, it would be of outmost importance to first systematically and objectively evaluate the strengths and weaknesses of the existing methods for different data sets, experimental conditions and questions, and only then start developing novel approaches for those settings that are beyond the currently available methods and tools. Therefore, after reviewing the existing imputation methods and their independent evaluation studies in the context of microarray data imputation, we gather some additional considerations and use cases that should be taken into account in the future, when designing more comprehensive evaluation frameworks. Besides helping the researchers to choose the imputation tool for a given data, such a practical evaluation framework should also benefit the developers of new imputation algorithms as a benchmark on which to test their new ideas in an objective and systematic manner. It is believed that there is still room for improvement in the field of missing value imputation, and new approaches are needed especially for non-microarray data sets, including those generated by mass-spectrometry-based proteomics, quantitative protein and genetic interactions screens, as well as by genome-wide association studies.

MISSING VALUE IMPUTATION IN GENE EXPRESSION MICROARRAY DATA

Although gene expression microarray platforms have developed significantly during the past years, the technology is still rather prone to errors, resulting in data sets with compromised accuracy and coverage. In particular, the presence of missing values due to various experimental reasons still remains a frequent problem, especially in complementary DNA (cDNA) microarray experiments. It has been shown in a number of studies that missing values in large-scale microarray data sets can drastically reduce their interpretation and hinder downstream analyses, including the performance of various downstream data analysis methods, such as unsupervised clustering of genes [6, 7], detection of differentially expressed genes [8, 9], supervised classification of clinical samples [10, 11] and construction of gene regulatory networks [12, 13]. To estimate the missing values before entering into such downstream analysis phases, several imputation tools have been developed for gene expression microarray data. In the following, the existing methods are roughly divided into two, partly overlapping categories: generic statistical methods and their application-specific modifications.

Generic statistical methods

The early strategies for imputing missing values in microarray data sets were highly inspired by the classic statistical models and estimation procedures. The aim of these strategies is to model the information captured in the non-missing part of the data set in order to estimate the missing values as accurately as possible. Some of the generally applicable principles for estimating missing data are as follows [1]: (i) 'mean imputation' is a simple procedure, in which the missing entries of the data matrix are estimated using the average of the non-missing values of the particular case or variable (row average or column average, respectively); (ii) 'hot deck imputation' involves predicting missing values using similar non-missing cases, where the neighbourhood can be defined using a distance function or metric (the so-called nearest-neighbours hot deck); (iii) 'model-based imputation' employs a statistical model, typically linear regression, to predict the missing values using the non-missing data of the same case, where the model can be estimated from cases with both observed and missing variables present (e.g. using the expectation-maximization (EM) algorithm [14]); (iv) 'multiple imputation' methods estimate more than one value for each missing entry, whereby the downstream methods can be applied to each complete data set individually, and these inferences can then be combined to produce the final result which also reflects the sampling variability due to missing values under the selected sampling scheme (e.g. Markov Chain Monte Carlo sampling); and (v) 'cold deck imputation' uses an external source of information, such as data from other similar studies, to estimate the missing values in the present study (compare with meta-analysis methods). Of course, 'composite methods' can also be defined that combine ideas from different approaches [1].

Perhaps the first imputation method specifically developed for microarray data was based on the K-nearest-neighbours version of the hot deck imputation, named KNN imputation, in which a set of K predictor genes with an expression profile similar to that of the gene with the missing value are first selected using Pearson's correlation or Euclidean distance, and then a distance-weighted average over the K genes is used as an estimate for the missing value [15]. It was shown that KNN imputation could drastically improve upon the simpler approaches, such as zero imputation or row average, and it has remained a popular option in many applications. Modifications to this generic procedure include, for instance, using a linear or non-linear regression models in place of the weighted average in the estimation step, and Bayesian variable selection or Spearman's correlation measure in the selection of the most predictive genes for imputation [16–20]. A set of methods based on the least squares principle, named LS imputation, utilizes correlations between both genes and arrays, and then allows combining of these estimates using an adaptive procedure [17]. A local version of the same approach, named LLS imputation, has gained much popularity, perhaps because its implementation also includes a procedure for choosing the neighbourhood size, K, automatically from the given data [20]. As an alternative to using directly the observed genes or arrays as predictors, a number of papers have proposed estimating the expression levels by means of singular value decomposition [15, 21], Gaussian mixture models [7] or support vector regression [22]. These global methods typically involve parameters analogous to the neighbourhood size that adjust the complexity of the predictive model. Some methods,

Table 1: Determinants of representative missing value imputation methods

Imputation model (abbreviated name)	Prediction variables	Estimation method*	Ref.
K-Nearest neighbours (KNN impute)	Matrix rows (genes)	WA	[15]
Least squares regression (LS impute)	Matrix rows or columns (arrays)	LS	[17]
Local least squares (LLS impute)	Matrix rows or columns	LS	[20]
Singular value decomposition (SVD impute)	Singular vectors ('eigengenes')	EM	[15]
Bayesian principal-component analysis (BPCA)	Principal components	EM	[23]
Gaussian mixture clustering (GMC impute)	Gaussian components	EM	[7]
Support vector regression (SVR impute)	Support vectors	QP	[22]

Most of the imputation procedures include at least the following three components: selection of a set of prediction variables to be used in the imputation; estimation of the predictive model; and finally, prediction of the missing values under the estimated model.
*WA, weighted average; LS, least-squares optimization, EM, expectation-maximization algorithm; QP, quadratic programming.

such as BPCA imputation, determine the number of prediction variables, such as the principal axes, automatically [23]. The basic concepts behind some of these methods are summarized in Table 1.

'Sequential imputation' refers to an implementation strategy, in which the imputation process is conducted successively in distinct data groups, arranged according to their missing value rate [1]. Several versions of such a strategy have been introduced also in the context of microarray data imputation; they all start the imputation process from the genes with the fewest missing values, and then sequentially utilize the imputed values in the later imputation steps using, for instance, KNN or LLS imputation [24–26]. The sequential approaches can also be coupled with an EM type of estimation [24], multiple imputation [25] or with an automated neighbourhood selection for each target gene separately [26]. 'Iterative imputation' is another widely used implementation strategy, which shares similarities with sequential imputation; it is actually an example of the EM algorithm, in which the imputation process starts with trial values substituted for all missing entries, and then the new estimates are iteratively updated on the basis of the current data, until they converge [1]. This strategy has been applied to microarray data imputation using row averages in the initialization step and either LLS or KNN in the iterated estimation steps [27, 28]. The iterated version of the local least squares method, named ILLS, also introduced a variant of the local neighbourhood adjustment, in which all of the genes that are within a distance threshold of the target gene are selected for the imputation [27]. In addition to these single imputation methods, there also exist versions of 'multiple imputation' for microarray data, in which several linear regression-based predictions are first calculated, and then an average over these is used as a final estimate for the missing value [24, 29, 30].

Application-specific modifications

The second generation of methods for imputing gene expression microarray data were built upon specific properties of the data sets being imputed, such as their quality issues or experimental designs. For instance, in the spotted cDNA arrays, poor-quality spots are often filtered out, either flagged up manually or identified through quality measures. Even though such filtering often constitutes a major reason for missing data, and may even lead to unwanted bias if the filtering process is highly selective in the types of genes affected, the spot quality measures were not explicitly used in the early missing value imputation approaches. At least two studies have addressed this issue; the first one uses posterior probabilities from a Gaussian mixture model to first decide whether imputation is required at all, and then to either impute the poor-quality spots or down-weight their contribution to the downstream analyses [31]. The second study incorporates a continuous spot quality weight directly into the estimation process, thereby allowing good-quality spots to have more impact on the imputation of other spots and to be themselves subject to less imputation than spots of a poorer quality [32]. Outlying observations, such as extreme expression levels, can also have severe effects on the imputation. To make imputation methods more robust against the outliers, one can employ quantile regression in place of the least squares optimization [33], or robust counterparts of the mean and covariance measures [34]. Finally, most of the generic imputation algorithms may not be particularly suitable for time-series experiments, especially when a particular time point contains many missing values. Imputation approaches tailored to the specific needs of the longitudinal experimental designs include, for instance, cubic splines [35], hidden Markov models [36], dynamic time warping [37], autoregressive models [38] and impulse models [39].

All of the imputation methods described so far estimate the missing expression values solely on the basis of the expression data. However, it has been postulated that auxiliary sources of information on the genes being imputed could help the estimation, in particular, when the primary expression data are somehow limited. The methods that use external information sources can be considered as versions of cold deck imputation, but they are reviewed in this subsection because such additional information must evidently be application specific to be useful for the imputation process. As far as we know, the first method that illustrated this idea took advantage of gene functional annotation information from gene ontology (GO) as a source of external data; more precisely, GO-based semantic similarity was combined with the expression similarity when selecting the relevant genes for imputation [40]. After the publication of this generic framework, other external information has also been shown to increase the accuracy of traditional methods, typically KNN and LLS, especially when the number of arrays is small or the proportion of missing values is high. A particularly useful source of information for predicting expression levels—as could be expected—originates from gene regulatory mechanisms in the form of, for instance, promoter sequence binding information on transcription factors [41], or histone acetylation state information on chromatin structure [42]. Alternatively, an extensive source of additional data comes from existing microarray studies that have studied similar or related experimental questions. Such meta-data-based methods for missing value imputation can utilize publicly available databases, such as ArrayExpress, Gene Expression Omnibus or Stanford Microarray Database, to improve the standard imputation methods, including KNN, LSS or Gaussian mixture clustering [43, 44]. As expected, the improvements are highest in the small sample sizes and noisy data settings, because the additional data sets can compensate up to a certain degree for the limited amounts and poor quality of the primary microarray data.

Although the imputation methods that make use of external information sources have shown improved performance, as compared to their counterparts that use expression data alone, they also come with some limitations. For instance, the meta-data-based methods are obviously not applicable in those cases where no related data sets exist for the particular experimental setting or study organism. This may especially be the case when non-standard experimental conditions or non-model study organisms are involved. Moreover, even if previous data exist, it may be difficult to find relevant data sets in the public microarray repositories, as many data sets are rather poorly annotated. The availability of external information is a major limitation also for other integrative imputation methods. For instance, transcription factor-binding information is mapped only very scarcely for many organisms, and it ignores other transcriptional regulation mechanisms, such as histone modifications. On the other hand, histone acetylation information is not available even for all of the genes in yeast, not to speak of other organisms or experimental conditions. Finally, while the GO annotations are available for most organisms, the GO-based imputation is subject to the accuracy and coverage of the gene functions annotated in the databases. However, with the increasing amount and accuracy of data sets and information accumulating in the public databases and information repositories, these methods are likely to become more useful in the future. In addition to the local imputation approaches, such as KNN and LSS, it would be interesting to combine additional information also into other, more sophisticated imputation approaches that consider also the global correlation structures in the data sets. Such future methods could be implemented using the computational framework, called projection on the convex sets, or POCS, in which several *a priori* characteristics of a data set, either internal or external, can be formulated as constrains in the missing value estimation problem [45].

EMPIRICAL GUIDELINES FOR CHOOSING AN IMPUTATION METHOD

Several recent publications have introduced new imputation algorithms for microarray data, and the authors of these algorithms have shown improved performance over the previous approaches in selected data sets that satisfy the specific assumptions of the algorithm. For instance, it has been long recognized that even the simplistic methods, such as row average, may provide sufficient estimates in data sets without an underlying correlation structure, but more sophisticated methods, such as KNN imputation, are needed when marked correlations between genes and/or arrays exist [15, 46]. Beyond these individual observations made over a rather limited number of data sets, there are only a few independent comparisons of the relative performance of the state-of-the-art imputation methods using a common and comprehensive set of microarray data. Recently, Brock *et al.* [47] carried out an extensive evaluation of eight current imputation methods on nine data sets of various sizes and type, including time series, multiple exposures and mixed type of data, to assess their performance under different conditions and establish guidelines for their appropriate use. The methods were compared at different percentages of missing data in terms of the similarity between the original and imputed data points; more formally, the imputation accuracy was assessed using the root mean squared error (RMSE) calculated on log-transformed expression data matrices. They found that the top-performing imputation algorithms, namely LS, LLS and BPCA, are all highly competitive with each other, but no method is uniformly superior in all of the data sets examined; global imputation approaches, such as BPCA, performed better on microarray data in which gene expression values are strongly correlated, whereas neighbour-based methods, such as LS and LLS, performed better in data with a local substructure. Accordingly, they proposed an entropy measure to quantify the complexity of expression matrices, which could be used when selecting an appropriate imputation algorithm for a given data set [47].

While most of the imputation algorithms have been evaluated only in terms of imputation accuracy, using metrics such as RMSE, it can be argued that the success of expression value estimation should be evaluated also in more practical terms; ideally, with respect to the downstream objective of the experiment, the motivation being that if the differences between the outcomes are biologically insignificant, then it is irrelevant whether the improvements at measurement level are statistically significant or not. Cluster analysis is typically one of the first downstream analyses conducted for a gene expression microarray data set since it provides, for instance, hypotheses about functional roles of the genes for subsequent experimental or computational analyses. However, it has been observed that even a small number of missing values may dramatically decrease the stability of clustering algorithms, such as hierarchical and *k*-means clustering, but their stability can be improved using imputation methods, such as KNN and GMC impute [6, 7]. Recently, Tuikkala *et al.* [48] carried out a systematic comparison of more recent missing value imputation methods by following the typical microarray data analysis work flow: eight recent microarray data sets, including both time series and steady-state experiments, were first clustered using the *k*-means algorithm, and then the resulting gene groups were interpreted in terms of their enriched GO annotations. They found out that even when there are marked differences in the measurement-level imputation accuracies across the imputation methods, as assessed using RMSE, these differences may become negligible when the methods are evaluated in terms of how well they can reproduce the original gene clusters or their biological interpretations. Regardless of the evaluation approach, however, these results strongly supported the earlier observations that more advanced imputation methods, such as BPCA, always provide much better downstream analysis results than ignoring the missing values or replacing them with zeros or average values [48].

Beyond the cluster-level evaluations, selected missing value imputation methods have also been compared with respect to other downstream objectives. For instance, it has been observed that missing values can drastically affect the detection of differential expression, and that the more sophisticated methods, such BPCA and GMC imputation, can much better handle the missing values when compared to the simple Row average or KNN methods [8]. Several studies have also investigated the effect of missing values on biomarker discovery and disease classification. In one study, it was concluded that while zero imputation resulted in poor classification accuracy, KNN, LLS and BPCA imputation methods had relatively minor differences with respect to the classification performance [10], whereas another study concluded that both the BPCA and ILLS methods could preserve the classification accuracy achieved on the complete data [11]. Further studies are needed to fully evaluate the relative performance of the different approaches and implementations. Although the running time of the implementations can be considered as a secondary evaluation criterion, it is still worth noting that BPCA imputation is one of the fastest methods among the more advanced imputation approaches, being only 10 times slower than the simple KNN, and that the running time of the LLS method is about one-third of that of its iterative version ILLS, which is one of the slowest imputation methods [48]. Besides being slow, the ILLS method produces, in some data sets, estimates for missing values that can be up to 10 times larger than the original values [48], which suggest an anomaly in its implementation or operation. Similarly, the LLS method was found to perform sub-optimally in some complex data sets, whereas the performance of the original least squares implementation, LS impute, was more consistent across the data sets [47]. This suggests some differences in the way the local regression problem is treated in the observed data matrix, which can lead to marked differences in some microarray data sets.

Taken together, these results demonstrate that the negative effects of missing values on the results of several downstream outcomes can—up to a certain degree—be attenuated using readily available and relatively fast and robust imputation methods. In particular, the BPCA method is available as Java, MATLAB and R implementations [49], and it is also integrated into an R-package, called arrayImpute, which implements a number of imputation algorithms, including KNN and LLS impute [50]. The original LS impute is implemented as a Java application [51].

ADDITIONAL RESEARCH TOPICS FOR DEALING WITH MISSING VALUES

Suggestions for future evaluation studies

Despite the few comparative studies that have systematically evaluated the relative merits of various state-of-the-art imputation methods using common data sets, it remains largely unclear which method is preferable for which experimental setting and question. This is because even the systematic and objective comparative studies have been rather limited in terms of the number of different data sets and evaluation criteria used in the comparisons. Consequently, there is only scattered information on practical questions, such as what are the specific advantages and potential limitations of the different approaches under different settings, or are there even any marked differences between the methods in more practical terms? In particular, the observed dependence of the RMSE on the data set can seriously bias the evaluation of imputation algorithms. In fact, it enables the authors to choose those data sets that favour their imputation method. This makes it difficult to grasp the overall picture of the strengths and weaknesses of different methods for different data sets. Independent, comprehensive evaluations are therefore warranted that systematically compare the existing methods over a large number of data sets that include experimental designs encountered in practice, such as time series, steady-state and mixed experiments, and hence allow one to investigate the dependence of the imputation results on various data properties, for instance, the dimensionality and correlation structure of the expression data matrix. Further, while the existing methods have mainly been evaluated in data sets from model organisms, typically yeast, it would be extremely informative to include in future comparisons also data from other organisms, including humans. Additional characteristics that should be considered include, for instance, the type of samples (e.g. controlled sample material or heterogeneous clinical specimens) and data pre-processing strategies used (e.g. gene filtering and data normalization).

Another potential pitfall in the current evaluations lies with the distribution of the missing entries. Most of the imputation methods have been developed and validated under the assumption that missing values occur completely at random within the expression data matrix. However, this assumption does not always hold in practice since the different measurements (arrays) are often conducted under variable experimental conditions, hence leading to differences in hybridization, media or time, among other factors. Consequently, the distribution of missing entries in microarray experiments can be highly non-random. This issue should be taken into consideration when interpreting the comparative evaluations or when designing fair evaluations and improved methods. For instance, if the missing value distribution in a comparison is forced to be the same as one used by a particular method, for instance, in parameter optimization, then it may lead to overoptimistic results. It has been observed that the imputation accuracies of LLS and ILLS, in particular, seem to benefit from the uniform missing value distribution [48]. Similarly, if a method uses the assumption of random missing values in the selection of appropriate imputation algorithms or their combination, when in fact this is not the case, it may lead to sub-optimal imputation results in practice. The mechanism by which the missing values are generated can have a marked effect on the imputation results. In particular, if the methods are evaluated solely in terms of RMSE, then one has to pay particular attention to the missing value distribution of the data sets. It is therefore recommended that the future comparative evaluations also take into consideration evaluation criteria other than the imputation accuracy or cluster stability. Additional outcomes, such as the detection of differentially expressed genes, biological significance of the detections or construction of gene regulatory networks, have already been investigated in recent works that introduce new imputation methods [12, 28, 30], but more comprehensive and objective future studies are needed, similar to the existing community efforts, such as AffyComp that benchmark the competing expression-level estimates [52].

Imputation tools for other large-scale data sets

The imputation methods described here can, in principle, be modified for other large-scale data sets that share similarities with microarray gene expression data. In particular, the recent advances in liquid chromatography-mass spectrometry (LC-MS)-based proteomic approaches are enabling quantitative profiling of complex peptide mixtures in sample sizes large enough to allow many interesting applications, including protein biomarker discovery by means of statistical testing, and clinical sample classification by means of machine learning methodologies. However, despite its tremendous potential, the methodological sophistication still lags well behind that of routine transcriptomic experiments. In particular, missing peptide identifications and quantifications remain a frequent problem, especially in large-scale clinical proteomic studies. Besides the frequent missing values, some other aspects of proteomic studies, such as the relatively small sample sizes compared to the complexity of the peptide mixtures, are similar to those in gene expression microarray studies. However, proteomic data also have more intrinsic properties; for instance, the missing peptide identifications may occur more frequently in patients than in control samples, due to true biological differences and technical bias in the MS instruments towards detecting high-abundance proteins. Therefore, the missing data should not be treated as missing completely at random, but as missing at random, at least in the case when the reason for missingness depends upon other observed patient characteristics. Accordingly, the missing value estimation methods developed for microarrays, such as KNN imputation that is being applied to statistical analysis of quantitative LC-MS-based proteomics data [53], may yield sub-optimal results; it is likely that more application-specific methods are needed that also account for the missingness mechanisms [54]. Moreover, it may be beneficial to incorporate other relevant data sources for missing value imputation, such as the clinical annotation of the specimens [55], or messenger RNA (mRNA) abundance, cellular role and other available information on the proteins not experimentally detected [56].

To promote their widespread usage in various large-scale data sets, the missing value imputation methods should be made publicly available, preferably as software packages with a graphical user interface (GUI), that are easily accessible to the experimental practitioners. A good example is the arrayImpute and arrayMissPattern software, which, in addition to incorporating a number of advanced imputation algorithms into its user-friendly GUI, provides several visualization options, for instance, to check whether or not the patterns of missing values occur at random, and comparison of the imputed values obtained from various methods to allow the user to experiment with different algorithms and parameter combinations [50]. It should be noted that methods that do not require parameter adjustment, such as LS and BPCA, are often attractive choices for the average users. Although some of the imputation algorithms, including BPCA, have been implemented as stand-alone software tools with a GUI [49], it would be valuable to include the state-of-the-art imputation methods in some popular data analysis packages as optional data pre-processing methods, alongside with data filtering and normalization. At the moment, however, packages such as SAM [57], Expression Profiler [58] and DAnTE [59], include only standard algorithms, such as KNN or SVD-based imputation, that may provide sub-optimal results in many applications. The advanced imputation methods should also benefit other microarray technologies, including exon, tissue and protein arrays, as well as whole-genome chromatin immunoprecipitation

(ChIP)-on-chip or single-nucleotide polymorphism (SNP) genotyping arrays. However, it is expected that novel imputation approaches are needed for many emerging data sets (with intrinsic properties), including those from the quantitative screening of protein–protein interactions (PPI; high dimensionality and noise levels, incomplete design matrices), or meta-analyses among genome-wide association studies (GWAS; high dimensionality and inter-individual variation, categorical value data points). It is also interesting to note the connection between missing value imputation and prioritizing of future screens, for instance, when completing genome-wide mappings of gene or protein interactions. As a full discussion of these applications is beyond the scope of this paper, interested readers are referred to some recent papers [60–65].

Methodologies that can cope with missing values

Imputation is a general and flexible method for handling problems posed by missing data values; however, it is not without pitfalls: ''The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases'' [1]. For instance, a particular challenge beyond the capacity of standard imputation algorithms is how to accurately estimate extreme expression levels in microarray or proteomic experiments. In two-dye microarray platforms, extremely low or high missing value ratios may arise, for instance, from an undetectable low signal in one of the channels only, and such intensity-dependent missing values have been shown to affect more severely the imputation results and downstream analyses than those whose values are independent—or near the centre—of the expression distribution [6, 9, 18, 66]. Similarly, in the LC-MS platforms, low-abundance proteins tend to remain undetected, due to the detection bias of the current MS-instruments, hence resulting in many intensity-dependent missing values, which may lead to the loss of potential protein biomarkers. Both of these examples present cases of 'missing not at random', because the reason for missingness is related to some unobserved data characteristic (here the expressions value itself, that is, the data is said to be 'censored'). This situation is different from a non-uniform distribution of missing entries within the data matrix, because the rows and columns with missing values can be observed directly from the data, and hence this case can still be treated as a normal missing at random situation. However, in the missing not at random cases, valuable information on the missingness mechanisms is lost, and therefore it may well be that no feasible imputation solutions for estimating such missing patterns will exist [1, 2].

Moreover, as the eventual goal of data mining is not to predict the expression values themselves, but rather to facilitate revealing the most important findings from the large-scale studies, methods that can handle the missing values as a part of their operation would be ideal for many real applications. For dealing with incomplete data matrices without the use of imputation, some of the general-purpose procedures, such as SVD and PCA, can be made robust to missing or outlying data values [67, 68]. There are also some classes of statistical procedures that are particularly well suited to handling missing values. In the likelihood-based statistical modelling, the EM algorithm or its modifications can be used to estimate the pre-specified model from incomplete data [1, 14]. When combined with mixture models, for instance, the EM algorithm can be used both for the estimation of mixture components and for coping with missing data, with applications to a wide range of supervised and non-supervised learning problems [69]. In particular, mixtures of hidden Markov models have recently been used for the robust classification of clinical time series with missing data points [70]. In Bayesian learning, missing values can be treated as additional unknown variables and iteratively imputed using, for instance, Gibbs sampling and data augmentation [71, 72]. Such a Bayesian approach with an additive model was recently proposed for bi-clustering of gene expression data with missing data [73]. There are also more straightforward iterative procedures that combine the gene expression clustering and missing value imputation by explicitly modifying the distance function [74, 75]. Finally, we and others have shown that some non-parametric procedures, based on bootstrapping and subsampling, can be surprisingly good at coping with missing data when detecting, for instance, differentially expressed genes or proteins, especially in small sample sizes, or hidden sample sub-groups that show coherent expression changes in clinical genome-wide studies [76–78].

CONCLUSION

Missing values remain a frequent problem in large-scale profiling experiments, such as those based on gene expression microarray or mass-spectrometry biotechnologies, and their adverse effect on the downstream analysis results is beyond the capacity of simple imputation methods, such as ignoring the missing data points or replacing them with zeros or average values. Although a wide spectrum of missing value imputation methods is available to the users of the high-throughput platforms, and new ones are constantly being developed, there is only limited guidance on how to choose between the different methods. The performance of the imputation methods may vary drastically depending on the experimental settings and questions under study. This discrepancy can partly be attributed to the way in which imputation methods have traditionally been developed and evaluated; in fact, with respect to some downstream objectives, such as gene clustering, the more sophisticated methods do not seem to differ much, even if there are marked differences in the measurement-level imputation accuracies across data sets. Therefore, instead of introducing incremental variations to the current imputation approaches, it would make an important and timely contribution to the community to systematically evaluate whether there are practical differences between the existing methods; and if there are, to pinpoint the preferable approaches under different platforms and experimental settings. This is important because otherwise the experimental practitioners will soon be overwhelmed by the huge amount of imputation tools without really knowing their real benefits and potential limitations. Besides helping the researchers to choose an appropriate imputation tool for a given data set, a systematic benchmarking framework would also benefit the developers of new imputation algorithms for data sets outside the scope of the existing approaches. Before such benchmarking results become available, it is recommended to rely on the robust imputation methods that have been shown to perform generally well under multiple settings, such as LS and BPCA; or alternatively, one can weight the imputation results obtained from several methods, using software tools such as ArrayImpute that allow experimenting with several imputation algorithms and with different parameter combinations.

Key Points

- The performance of a missing value imputation method depends on the properties of the datas et being imputed and on the experimental question under study.
- Further comparative studies that systematically evaluate the methods under different settings are warranted to provide practical guidance to the practitioners.
- Novel imputation approaches and methods that can handle missing values are likely to be needed in the coming days, especially for non-microarray data sets.

FUNDING

Academy of Finland (grant number 120 569).

References

1. Little RJA, Rubin DB. *Statistical Analysis with Missing Data.* New York: John Wiley & Sons, Inc, 1987.
2. Donders AR, van der Heijden GJ, Stijnen T, *etal*. Review – a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;59:1087–91.
3. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997;278:680–6.
4. 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;9:3273–97.
5. Alizadeh AA, Eisen MB, Davis RE, *et al*. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000;403:503–11.
6. 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.
7. Ouyang M, Welsh WJ, Georgopoulos P. Gaussian mixture clustering and imputation of microarray data. *Bioinformatics* 2004;20:917–23.
8. Jörnsten R, Wang HY, Welsh WJ, *et al*. DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics* 2005;21:4155–61.
9. Scheel I, Aldrin M, Glad IK, *etal*. The influence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics* 2005;21:4272–9.
10. Wang D, Lv Y, Guo Z, *etal*. Effects of replacing the unreliable cDNA microarray measurements on the disease classification based on gene expression profiles and functional modules. *Bioinformatics* 2006;22:2883–9.
11. Shi Y, Cai Z, Lin G. Classification accuracy based microarray missing value imputation. In: Mandoiu I, Zelikovsky A, (eds). *Bioinformatics Algorithms: Techniques and Applications*. NJ: Wiley-Interscience, 2007;303–28.
12. Sehgal MS, Gondal I, Dooley LS, *et al*. How to improve postgenomic knowledge discovery using imputation. *EURASIP J Bioinform Syst Biol* 2009, Article ID 717136.

13. Zhang Y, Xuan J, de los Reyes BG, *et al.* Reverse engineering module networks by PSO-RNN hybrid modeling. *BMC Genomics* 2009;10(Suppl 1):S15.
14. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc B* 1977;39:1–38.
15. Troyanskaya O, Cantor M, Sherlock G, *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001;17:520–5.
16. Zhou X, Wang X, Dougherty ER. Missing-value estimation using linear and non-linear regression with Bayesian gene selection. *Bioinformatics* 2003;19:2302–7.
17. Bø TH, Dysvik B, Jonassen I. LSimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Res* 2004;32:e34.
18. Nguyen DV, Wang N, Carroll RJ. Evaluation of missing value estimation for microarray data. *J Data Sci* 2004;2: 347–70.
19. Brás LP, Menezes JC. Dealing with gene expression missing data. *Syst Biol* 2006;153:105–19.
20. Kim H, Golub GH, Park H. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics* 2005;21:187–98.
21. Friedland S, Niknejad A, Chihara L. A simultaneous reconstruction of missing data in DNA microarray. *Linear Algebra Appl* 2006;416:8–28.
22. Wang X, Li A, Jiang Z, Feng H. Missing value estimation for DNA microarray gene expression data by support vector regression imputation and orthogonal coding scheme. *BMC Bioinformatics* 2006;7:32.
23. Oba S, Sato MA, Takemasa I, *et al.* A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* 2003;19:2088–96.
24. Kim KY, Kim BJ, Yi GS. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinformatics* 2004;5:160.
25. Verboven S, Branden KV, Goos P. Sequential imputation for missing values. *Comput Biol Chem* 2007;31: 320–7.
26. Zhang X, Song X, Wang H, *et al.* Sequential local least squares imputation estimating missing value of microarray data. *Comput Biol Med* 2008;38:1112–20.
27. Cai Z, Heydari M, Lin G. Iterated local least squares microarray missing value imputation. *J Bioinform Comput Biol* 2006; 4:935–57.
28. Brás LP, Menezes JC. Improving cluster-based missing value estimation of DNA microarray data. *Biomol Eng* 2007;24: 273–82.
29. Sehgal MS, Gondal I, Dooley LS. Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. *Bioinformatics* 2005;21: 2417–23.
30. Sehgal MS, Gondal I, Dooley LS, *et al.* Ameliorative missing value imputation for robust biological knowledge inference. *J Biomed Inform* 2008;41:499–514.
31. Tom BD, Gilks WR, Brooke-Powell ET, *et al.* Quality determination and the repair of poor quality spots in array experiments. *BMC Bioinformatics* 2005;6:234.
32. Johansson P, Häkkinen J. Improving missing value imputation of microarray data by using spot quality weights. *BMC Bioinformatics* 2006;7:306.
33. Yoon D, Lee EK, Park T. Robust imputation method for missing values in microarray data. *BMC Bioinformatics* 2007; 8(Suppl 2):S6.
34. Branden KV, Verboven S. Robust data imputation. *Comput Biol Chem* 2009;33:7–13.
35. Bar-Joseph Z, Gerber GK, Gifford DK, *et al.* Continuous representations of time-series gene expression data. *J Comput Biol* 2003;10:341–56.
36. Schliep A, Schönhuth A, Steinhoff C. Using hidden Markov models to analyze gene expression time course data. *Bioinformatics* 2003;19(Suppl 1):i255–63.
37. Tsiporkova E, Boeva V. Two-pass imputation algorithm for missing value estimation in gene expression time series. *J Bioinform Comput Biol* 2007;5:1005–22.
38. Choong MK, Charbit M, Yan H. Autoregressive-model-based missing value estimation for DNA microarray time series data. *IEEE Trans Inf Technol Biomed* 2009;13:131–7.
39. Chechik G, Koller D. Timing of gene expression responses to environmental changes. *J Comput Biol* 2009;16:279–90.
40. Tuikkala J, Elo L, Nevalainen OS, *et al.* Improving missing value estimation in microarray data with gene ontology. *Bioinformatics* 2006;22:566–72.
41. Elo LL, Tuikkala J, Nevalainen OS, *et al.* Predicting gene expression from combined expression and promoter profile similarity with application to missing value imputation. In: Deutsch A, Brusch L, Byrne H, de Vries G, Herzel H, (eds). *Mathematical Modeling of Biological Systems*; Vol. I. Boston: Springer- Birkhäuser, 2007.
42. 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.
43. Hu J, Li H, Waterman MS, *et al.* Integrative missing value estimation for microarray data. *BMC Bioinformatics* 2006;7: 449.
44. Jörnsten R, Ouyang M, Wang HY. A meta-data based method for DNA microarray imputation. *BMC Bioinformatics* 2007;8:109.
45. Gan X, Liew AW, Yan H. Microarray missing data imputation based on a set theoretic framework and biological knowledge. *Nucleic Acids Res* 2006;34:1608–19.
46. Feten G, Almøy T, Aastveit AH. Prediction of missing values in microarray and use of mixed models to evaluate the predictors. *Stat Appl Genet Mol Biol* 2005;4, Article10.
47. 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.
48. 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.
49. BPCA missing value tools. http://hawaii.aist-nara.ac.jp/~shige-o/tools/BPCAFill.html (30 September 2009, date last accessed).
50. Lee EK, Yoon D, Park T. arrayImpute – software for exploratory analysis and imputation of missing values for microarray data. *Genomics Informatics* 2007;**5**:129–32. http://bibs.snu.ac.kr/software/arrayImpute (27 October 2009, date last accessed).
51. LSimpute software supplementary web page. http://www.ii.uib.no/~trondb/imputation (27 October 2009, date last accessed).

52. Irizarry RA, Wu Z, Jaffee HA. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* 2006;22: 789–94.

53. Roxas BA, Li Q. Significance analysis of microarray for relative quantitation of LC/MS data in proteomics. *BMC Bioinformatics* 2008;9:187.

54. Karpievitch Y, Stanley J, Taverner T, *et al*. A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics* 2009;25:2028–34.

55. Emerson JW, Dolled-Filhart M, Harris L, *et al*. Quantitative assessment of tissue biomarkers and construction of a model to predict outcome in breast cancer using multiple imputation. *Cancer Inform* 2009;7:29–40.

56. Torres-García W, Zhang W, Runger GC, *et al*. Integrative analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: a non-linear model to predict abundance of undetected proteins. *Bioinformatics* 2009;25: 1905–14.

57. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001;98:5116–21.

58. Kapushesky M, Kemmeren P, Culhane AC, *et al*. Expression Profiler: next generation–an online platform for analysis of microarray data. *Nucleic Acids Res* 2004;32: W465–70.

59. Polpitiya AD, Qian WJ, Jaitly N, *et al*. DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics* 2008;24:1556–8.

60. Shoemaker BA, Panchenko AR. Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput Biol* 2007;3:e43.

61. Aryee MJ, Quackenbush J. An optimized predictive strategy for interactome mapping. *J Proteome Res* 2008;7:4089–94.

62. Schwartz AS, Yu J, Gardenour KR, *et al*. Cost-effective strategies for completing the interactome. *Nat Methods* 2009;6:55–61.

63. de Bakker PI, Ferreira MA, Jia X, *et al*. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 2008;17:R122–8.

64. Halperin E, Stephan DA. SNP imputation in association studies. *Nat Biotechnol* 2009;27:349–51.

65. Nothnagel M, Ellinghaus D, Schreiber S, *et al*. A comprehensive evaluation of SNP genotype imputation. *Hum Genet* 2009;125:163–71.

66. Ritz C, Edén P. Accounting for one-channel depletion improves missing value imputation in 2-dye microarray data. *BMC Genomics* 2008;9:25.

67. Liu L, Hawkins DM, Ghosh S, *et al*. Robust singular value decomposition analysis of microarray data. *Proc Natl Acad Sci USA* 2003;100:13167–72.

68. Stacklies W, Redestig H, Scholz M, *et al*. pcaMethods–a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* 2007;23:1164–7.

69. Ghahramani Z, Jordan MI. Supervised learning from incomplete data via an EM approach. In: Cowan JD, Tesauro G, Alspector J, (eds). *Advances in Neural Information Processing Systems*; Vol. 6. San Francisco, CA: Morgan Kaufmann Publishers, 1994.

70. Costa IG, Schönhuth A, Hafemeister C, *et al*. Constrained mixture estimation for analysis and robust classification of clinical time series. *Bioinformatics* 2009; 25:i6–14.

71. Geman S, Geman D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans Pattern Anal Machine Intel* 1984;12:609–628.

72. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. *J Amer Statist Assoc* 1987; 82:528–49.

73. Gu J, Liu JS. Bayesian biclustering of gene expression data. *BMC Genomics* 2008;9(Suppl 1):S4.

74. Kim DW, Lee KY, Lee KH, *et al*. Towards clustering of incomplete microarray data without the use of imputation. *Bioinformatics* 2007;23:107–13.

75. Wong DS, Wong FK, Wood GR. A multi-stage approach to clustering and imputation of gene expression profiles. *Bioinformatics* 2007;23:998–1005.

76. Hua D, Lai Y. An ensemble approach to microarray data-based gene prioritization after missing value imputation. *Bioinformatics* 2007;23:747–54.

77. Elo LL, Hiissa J, Tuimala J, *et al*. Optimized detection of differential expression in global profiling experiments: case studies in clinical transcriptomic and quantitative proteomic datasets. *Brief Bioinform* 2009;10:547–55.

78. Hiissa J, Elo LL, Huhtinen K, *et al*. Resampling reveals sample-Level differential expression in clinical genome-wide studies. *OMICS – J Integr Biol* 2009;13: 381–96.