Accounting for the multiple natures of missing values in label-free quantitative proteomics datasets to compare imputation strategies

Cosmin Lazar[1,3,4], Laurent Gatto[5,6], Myriam Ferro[1,3,4],
Christophe Bruley[1,3,4], Thomas Burger[1,2,3,4,*]

[1] Univ. Grenoble Alpes, iRTSV-BGE, F-38000 Grenoble, France.
[2] CNRS, iRTSV-BGE, F-38000 Grenoble, France.
[3] CEA, iRTSV-BGE, F-38000 Grenoble, France.
[4] INSERM, BGE, F-38000 Grenoble, France.
[5] Computational Proteomics Unit, Cambridge, CB2 1GA, UK.
[6] Cambridge Center for Proteomics, Cambridge, CB2 1GA, UK.
[*] thomas.burger@cea.fr

February 18, 2016

Abstract: Missing values are a genuine issue in label-free quantitative proteomics. Recent works have surveyed the different statistical methods to conduct imputation and have compared them on real or simulated datasets, and recommended a list of missing value imputation methods for proteomics application. Although insightful, these comparisons do not account for two important facts: (*i*) depending on the proteomics dataset, the missingness mechanism may be of different natures, and (*ii*) each imputation method is devoted to a specific type of missingness mechanism. As a result, we believe that the question at stake is not to find the most accurate imputation method in general, but instead, the most appropriate one. In this article, we describe a series of comparisons that support our views: for instance, we show that a supposedly "under-performing" method (i.e. giving baseline average results), if applied at the "appropriate" time in the data processing pipeline (before or after peptide aggregation) on a dataset with the "appropriate" nature of missing values, can outperform a blindly applied, supposedly "better performing" method (i.e. the reference method from the state-of-the-art). This leads us to formulate few practical guidelines, regarding the choice and the application of an imputation method in a proteomics context.

Keywords: label-free relative quantitative proteomics; missing value imputation.

1 Introduction

The high rate of missing values in label-free quantitative proteomics is a major concern [1]. From the literature, in the case of LC-MS/MS approaches, it frequently ranges between $10 - 50\%$, while the proportion of peptides/proteins that exhibit at least one missing value can very high, ranging in between $70 - 90\%$ [2]. As a consequence, it was originally proposed to apply imputation methods originally developed for transcriptomics and microarray data analysis [3] to proteomics data. Then, more general methods, developed in a theoretical statistical context, were considered [4], and adapted to some extent to proteomics datasets [5]. To date, numerous methods exists and are available to any practitioner, either as independent packages [6, 7, 8], or through dedicated pipeline packages such as `MSnbase` [9]. In addition, several methods have been reported that successfully leverage on a multi-omics context to impute proteomics missing values on the basis of transcriptomics observed values [10, 11, 12]. Recently, a comprehensive survey [13] compared and discussed some well-known imputation algorithms in the context of proteomics applications. There are numerous conclusions that can be drawn from this survey, or from references therein.

First, there are multiple reasons why values are missing, accounting for biochemical and analytical (miscleavage, dynamic range, ionization competition, ion suppression, etc.) to bioinformatics mechanisms (peptide misidentification, ambiguous matching of the precursors in the quantitation step, etc.). However, regardless of their origins, missing values can be cast in three categories with regards to the statistical mechanisms that best describe them. In fact, statisticians have defined three types of missing values [4]:

- *Missing Completely At Random* (MCAR), which in a proteomics dataset, correspond to the combination and propagation of multiple minor errors or stochastic fluctuations (for instance, a miss-identified peptide can or cannot be balanced by the alignment of the precursor maps, leading to an abundance value, or on the contrary to a missing value). As a result, each missing value cannot be directly explained by the nature of the peptide, nor by its measured intensity [5]. As a result, MCAR affect the entire dataset with a uniform distribution.

- *Missing At Random* (MAR), which is a more general class than MCAR, where conditional dependencies are accounted for. In a proteomics dataset, it is classically assumed that all the MAR values are also MCAR, so that one is little interested in MAR [5]. However, some MAR imputation methods can also be used for MCAR missing values, and thus applied to proteomics datasets.

- *Missing Not At Random* (MNAR), which, on the contrary, have a targeted effect. In mass spectrometry-based analysis, chemical species whose abundance are close to the limit of detection of the instrument enough record a higher rate of missing values. This is why, MNAR-devoted imputation methods used in proteomics focus on left-censored data (that is, the distribution of which with respect to the abundance is truncated on the left side, *i.e.* on the region depicting the lower abundances).

Second, the statistics literature contains numerous imputation methods devoted to MCAR or MAR, while very few are devoted to MNAR. The reason for this asymmetry is simple: most of the MCAR/MAR mechanisms are generic to numerous application fields, so that it naturally focused statisticians' efforts. On the other hand, MNAR (including left-censored) mechanisms are discipline-specific, so that a precise understanding of the mechanism underlying the data generation is mandatory. This is why, in the comparisons depicted in [13], among the nine methods, only three MNAR-devoted approaches were considered, among which two are based on the same principle. Nonetheless, these nine methods have been compared on various datasets,

that are reported to have both MNAR and MCAR, yet in unknown proportions. As a result, even if a couple of MCAR/MAR devoted methods are shown to perform slightly better, it makes sense to wonder if this holds in general, or if it is dataset dependent.

Even though most of the conclusions of [13] are well supported, there is a need to consider the proportions of MCAR and MNAR as hidden variables. This idea is not new: several recent works have proposed to perform imputation by estimating models (with maximum-likelihood [14, 15] or with empirical Bayesian [16] methods) which are rich enough to account for both types of missingness mechanisms. To the best of our knowledge, no study has evaluated the behaviour of an imputation method devoted to MNAR (respectively devoted to MAR/MCAR) on a dataset containing mainly MCAR (respectively MNAR). However, this question is of prime importance to the practitioner, as it helps guiding the selection of an imputation algorithm according to the risk of corrupting the downstream analysis when using an unadapted imputation method.

In this work, we have considered real and simulated datasets on which MCAR and MNAR were introduced in controlled proportions and have compared the performances of various imputation methods. Numerous conclusions and recommendations can be drawn from these experiments. However, beyond them, our work pinpoints the fact that most of the conclusions regarding imputation methods cannot be claimed to hold in general. On the contrary, they should be contextualized according to each dataset, the proportion of missing values, and their nature.

2 Material

Simulated quantitative dataset

To generate artificial peptide abundance data, we used a simplified version of the model proposed in [5], which reads:

$$y_{ij} = P_i + G_{ik} + \epsilon_{ij} \tag{1}$$

where y_{ij} is the log-transformed abundance of peptide i in the jth sample, P_i is the mean value of peptide i, G_{ik} is the mean differences between the condition groups, and ϵ_{ij} is the random error terms which stands for the peptide-wise variance. Here, P_i is randomly generated from a Gaussian distribution with mean μ and standard deviation σ. The dynamic range of peptides (in logarithm scale) can be therefore approximated by $[\mu-3\sigma, \mu+3\sigma]$. We considered two groups k_1 and k_2 of replicates, for which P_i generation was conducted with $\mu = 1.5$ and $\sigma = 0.5$. For each of the two groups, we selected two disjoint subsets of peptides (20% of the total number of peptides) and we added G_{ik} randomly drawn from the distribution mentioned above, to simulate a differential abundance between the peptides. Finally, the random error term has also been simulated by random draws from a Gaussian distribution with zero mean and standard deviation $\sigma_\epsilon = 0.5$. With these parameters, we simulated a log-transformed peptide abundance table with $m = 1000$ peptides and $n = 20$ replicates (equally split into groups k_1 and k_2).

To derive the protein abundance data, a map describing the peptide/protein relationships has been randomly generated by randomly drawing m integers from $[1, m_{prot}]$ where m is the number of peptides and $m_{prot} < m$ is the number of proteins (m_{prot} was set to $m/2$).

Real quantitative dataset

As a complement to the simulated data, we considered a real and publicly available dataset, that has been collected during a study designed to compare human primary tumour-derived xenograph proteomes of the two major histological non-small cell lung cancer subtypes, adenocarcinoma (ADC) and squamous cell carcinoma (SCC), using Super-SILAC and label-free quantification [17]. The raw files were analyzed by MaxQuant (version 1.3.0.5). Peaks were

searched against the UniProt human database (released July, 2012; http://www.uniprot.org) using the Andromeda search engine included in MaxQuant. The dataset within this package contains proteins intensity for 6 ADC and 6 SCC samples. The complete MaxQuant output file is available on the repository of the ProteomeXchange Consortium [18], with the dataset identifier PXD000438.

As this study requires precisely controlling each missing values, one must work on a *complete dataset*, i.e. where no missing value shows up. This has been obtained from the raw peptide-level PXD000438 dataset by filtering out the peptides which contain at least one missing value. Finally, the complete peptide-level matrix was log-transformed and median normalized.

MCAR and MNAR incorporation

Let α and β be the rate of missing values and the MNAR ratio, respectively. They read:

$$\alpha = \frac{100 \cdot (\#\text{MNAR} + \#\text{MCAR})}{nm} \qquad \beta = \frac{100 \cdot \#\text{MNAR}}{\#\text{MNAR} + \#\text{MCAR}} \tag{2}$$

For a given combination of α and β, the missing values are incorporated in a complete dataset as follow:

MNAR values are incorporated using a stochastic threshold, as follows: one randomly generates a threshold matrix T from a Gaussian distribution with parameters $(\mu_t = q, \sigma_t = 0.01)$, where q is the α^{th} quantile of the abundance distribution in the complete quantitative dataset. Then, each cell (i, j) of the complete quantitative dataset is compared to $T_{i,j}$. If it is greater than or equal to $T_{i,j}$, the abundance is not censored. On the contrary, if it is strictly smaller than $T_{i,j}$, a Bernoulli draw with probability of success $\frac{\beta \cdot \alpha}{100}$ determines if the abundance value is censored (success), or not (failure).

MCAR values are incorporated by replacing with a missing value the abundance value of $nm\frac{(100-\beta)\alpha}{100}$ randomly chosen cells in the table of the quantitative dataset.

This strategy is summarized in Figure 1. We used it for any combination of values for $\alpha \in [2\%, 52\%]$ and $\beta \in [0\%, 100\%]$.

[Figure 1 about here.]

3 Methods

Imputation algorithms

Since an exhaustive comparison of the missing value imputation algorithms is beyond the scope of this study, we selected a set of characteristic and widely applied methods, representing different families of imputation procedures, and which are conceptually different. We considered:

- k**NN** (k Nearest Neighbours) [3]: for a peptide showing missing values, the method consists in: (i) Finding k most similar peptides to the one considered (using a particular distance measure, e.g. Euclidean distance of Pearson's correlation coefficient); (ii) Imputing each missing value by averaging the k peptide values from the same replicate where that missing value occurred. Preliminary exploration of the range of parameter k showed that the imputation accuracy was rather stable for any $k \in [10, 20]$, and reach its maximum to 11, so that we used this latter value.

- **SVDimpute** (Imputation with Singular Value Decomposition) [3]: The quantitative dataset is considered a matrix on which mean centering and k-rank SVD are iteratively applied (where $k \in [1, n/2]$ where $n/2$ is the number of replicates in a given condition group), up to some convergence criterion. In our case, k was tuned to 1 ($k = 1$ and $k = 2$ gave the greatest performances according to preliminary tests).

- **MLE** (Imputation based on Maximum Likelihood Estimation): Assuming the quantitative dataset obeys some law f_θ of unknown parameter θ, maximum likelihood estimation principle is used to derive an estimator $\hat{\theta}$ of θ, and missing values are then imputed by random draws of $f_{\hat{\theta}}$. The literature dedicated to missing value imputation based on MLE is vast, and we recommend [19, 20] for a comprehensive survey of the topic. In this work, we employed the implementation available in the `R` package `norm` [21].

- **MinDet** (Deterministic minimum imputation) [22, 23]: It simply replaces the missing values by the minimum value, either globally observed in the dataset, or observed in each sample. Here, we used the 10^{-4} quantile.

- **MinProb** (Probabilistic minimum imputation): It is a stochastic version of MinDet, so as to limit the bias introduced by multiple replacements with a unique value. The imputation is performed by replacing the missing values with random draws from a Gaussian distribution centered on the value used with MinDet, and with a variance tuned to the median of the peptide-wise estimated variances [24].

We decided to focus on these five methods, as they represent well the various types of imputation methods: First, according to the taxonomies provided in [25, 26], kNN, MinDet and MinProb belong to the *prediction rules* methods, SVDimpute belongs to the *least-square-based* methods, and finally, MLE belongs to the *maximum-likelihood-based* methods; so that, this set of methods covers well the taxonomies of [25, 26]. Second, according to [13], MinDet and MinProb are *single value approaches*, kNN is a *local similarity approach*, and SVDimpute is a *global similarity approach*; so that the taxonomy of [13] is also covered. Third, MinDet and MinProb are designed to impute MNAR values, while kNN, SVDimpute and MLE are designed for MCAR (and more generally MAR) values. Finally, MinDet is the most naive method to deal with MNAR (and often implemented as zero value imputation), while MLE and SVDimpute are particularly efficient on MCAR, so that comparing these three methods is insightful with regard to the conclusions of [13] on the general dominance of MCAR/MAR-devoted methods. Let us also notice that no multiple imputation method is considered in our work, while in practice, they provide the best results in the state-of-the-art. The reason is the following: Multiple imputation strategies amount to a boosting strategy, i.e. the combination of several simple methods to stabilize the results. However, their behavior, efficiency and adequation to the specificities of the data are directly related to those of the simple methods they are based of. As a result, we found it clearer to focus on the single imputation methods, so as to best describe and understand them, and to let the practitioner generalize our conclusion to multiple imputations. Finally, this set of algorithms has been chosen to represent a wide diversity of strategies, on which very general conclusions can be drawn.

Accuracy measurements

In most of the experiments, the imputation step was followed by the aggregation of peptide abundances into protein abundances (we estimated each protein abundance with the median abundance over the protein specific peptides). However, in few specific experiments (see Section 4), the aggregation was conducted first (i.e. on peptide abundances that still contain missing values), and followed by imputation at protein-level.

In both cases, we evaluated the performances of the imputation algorithms in the same way: we considered the differences between the protein abundances in the original complete quantitative dataset, and in its counterpart containing missing values that have been imputed (either at protein or peptide-level). Such differences are classically summarized by the *root mean square error* (RMSE), yet many other variants exist [27]. Within our framework, we employed a normalized version of the RMSE called the RMSE-observations standard deviation ratio (RSR) [28], defined as follows:

$$RSR(X_C, X_I) = \frac{RMSE(X_C, X_I)}{sd(X_C)} \tag{3}$$

where X_C denotes the complete quantitative dataset (before incorporating missing values), while X_I denotes the quantitative dataset after the imputation of the missing values. The reported results corresponds to an average over 30 independent repetitions of the experiment (i.e. the random generation of missing values as well as their imputation, for a given tuning of α and β), so as to have more stable performance records.

4 Results

MCAR-devoted vs. MNAR-devoted imputations

[Figure 2 about here.]

[Figure 3 about here.]

Figures 2 and 3 display a series of heatmaps (with a false color code, ranging from blue, which indicates low RSR, to red, which indicates high RSR) for the simulated and real datasets, respectively. Within each figure, there are five graphics, corresponding to an imputation method each. Each heatmap displays the average performances (over 30 repetitions) of the imputation algorithm over all the range of the experimental conditions (i.e. a proportion of missing values ranging from 2 to 52%, and an MNAR ratio ranging from 0 to 100%). Several conclusions can be drawn from these figures.

First, irrespective of the dataset, all methods perform better when there are less missing values, and become inaccurate with increasing proportion of missing values. Although expected, this result assesses the validity of our comparison protocol and of our simulations.

Second, two groups of algorithms can be identified, with regard to the MNAR ratio: the first group is made of SVDimpute, kNN and MLE, which perform better under a small MNAR ratio, while the second group, composed of MinDet and MinProb, performs better under a larger MNAR ratio. This clearly indicates that, depending on the nature of the majority of the missing values, it is important to privilege either a MCAR/MAR-devoted method, such as advocated in [13], or, on the contrary, to favour a MNAR-devoted method, even if the latter is more naive and provide, on average, worse results.

Third, for each method, a similar behaviour is observed on both the real and the simulated datasets. In the case of MinDet and MinProb, the similarity is almost perfect, with particular poor performance toward high percentages of non-random missing vales (lower right corner). In the case of the three other methods, even if the similarity between the heatmaps derived from the real and simulated datasets is not as good, a pattern is well-conserved. In both cases, the best performance is reached with the lowest rate of missing value and the lowest MNAR ratio (lower left corner), while the worst performance is reached with the greatest rate of missing value and the greatest MNAR ratio (upper right corner). In addition, isoperformance lines are roughly parallel to an axis going from the upper left to the lower right corner. The global stability of this pattern indicates that, even if MCAR is possibly a simplistic process to account for the diverse

nature of missing values that are not left-censored, the postulate at the root of these experiments is robust. Indeed, we postulated the strong influences of both (1) the rate of missing values and of the MNAR ratio, as well as (2) the nature of the missing values to which a given imputation method is devoted.

Finally, if one averages the performances of the various imputation methods over all the experiments (which amounts to consider a mean color over each graphic), it appears, that overall, MCAR/MAR-devoted methods (SVDimpute, kNN and MLE) outperforms MNAR methods (MinDet and MinProb). From this, we conclude that in absence of any knowledge regarding the MNAR ratio (and assuming that all the situations are equiprobable, which remains to be proven), it makes sense to privilege the former ones, such as advocated in [13].

However this averaging must not be overstated, as it is possible to show situations where even the worst MNAR method (MinDet) significantly outperforms the best MCAR/MAR methods (MLE or SVDimpute). To further demonstrate this, we applied an unpaired two sample t-test to assess the significance of the difference of accuracy, between the two following pairs of imputation methods: MinDet vs. SVDimpute (for the simulated dataset), and MinDet vs. MLE (for the real dataset). The results are reported in Figures 4. These comparisons demonstrate that when a high proportion (70% or more) of missing values are MNAR, MNAR imputation methods are preferred. Although such datasets are not widespread, they are not unheard of (see for instance [29, 30]), which advocates for the development of new methodologies that can estimate the nature of the majority of the missing values, so as to adapt the imputation method accordingly.

[Figure 4 about here.]

Peptide-level vs. protein-level imputations

In the literature, there is no consensus on the preferred order with respect to missing values imputation and aggregation of peptide intensities into protein intensities. This is why, both cases were considered in [13]. We also repeated our experiments summarized in Figures 2 and 3, in a reversed context where the aggregation is performed first, and the imputation is conducted at protein level. We have compared these two approaches using the methodology described above and present the results of a significance analysis (at a p-value threshold of 5%) in Figures 5 and 6, where blue indicates peptide imputation superiority, red indicates protein imputation superiority, and green indicates a non-significant result.

[Figure 5 about here.]

[Figure 6 about here.]

As illustrated by a high proportion of blue, peptide-level imputation is most of the time more accurate. Nevertheless, a major argument for protein-level imputation is the presence of less missing values; indeed, if several peptides are aggregated into a protein, this aggregation does not lead to a missing value, unless all the peptide intensities are missing, so that numerous missing values are implicitly imputed by a value which is a neutral element with respect to the aggregation. For instance, in the case where:

- protein intensities result from the sum of the peptide intensities (such as in [31]); then, missing peptide intensities do not contribute to the sum, so that the result is the same as if peptide missing values were imputed by zero.
- protein intensities result from the mean of the peptide intensities (such as in [32]); then, missing peptide intensities do not contribute to the mean, so that the result is the same as if peptide missing values were imputed by the mean value of the peptide intensities.

- protein intensities result from a maximum function of the peptide intensities (sum or mean over the three most abundant peptides, maximum peptide abundance, etc., such as in [33, 34]); then, the result is the same as if peptide missing values were imputed by zero, or any other small intensity.

For more general protein aggregation methods, based on more sophisticated functions (such as for instance, weighted mean), the issue is the same (even if the formula of the neutral element may be less trivial). The above observations are schematically described in Figure 7, where protein-level imputation is equivalent to (*i*) applying an implicit imputation method on some peptide-level missing values, that is neither controlled nor evaluated; (*ii*) performing the aggregation itself; (*iii*) explicitly imputing the few remaining missing values. As the total number of imputed missing values (whether implicit or explicit) is the same, it is preferable to consider an explicit and well-justified imputation for all the missing values, which amounts to impute at peptide-level and concurs with the results of Figures 5 and 6.

[Figure 7 about here.]

However, from Figures 5 and 6, it seemingly appears that when the data contain up to about 60% of MNAR values, and if an MNAR-devoted imputation method has been chosen *a priori*, it is more efficient to impute at the protein-level. This observation highlights that, on MCAR data, an implicit and sub-optimal imputation is more efficient than an MNAR imputation method. Deriving this result on the basis of the aforementioned observation (Figures 5 and 6) requires several steps:

1. During the aggregation process, several MCAR peptides are combined with observed peptide intensities (there is very little chance that, assuming MCAR data, all the peptides of a given protein are missing), leading to protein intensities rather than missing values.

2. As opposed to 1., let us note that MNAR peptides correspond to genuine low abundance ions, so that there are good chances that one aggregates only missing values, leading to a missing value at the protein level.

3. As a result from 1., it appears that if one has chosen to use a MNAR-devoted method, MCAR are either imputed by an unadapted method (at the peptide level), or implicitly imputed by the aggregation.

4. As a result from 2., if one uses the same MNAR-devoted method, MNAR are roughly imputed in the same way, both at peptide and at protein levels.

5. As a result from 3. and 4., one derives that the difference in the overall quality of the imputation (between peptide level and protein level imputation with an MNAR-devoted method) mainly relies on that of MCAR data.

6. Let us now recall the original observation: "when the data contain up to about 60% of MNAR values, and if an MNAR-devoted imputation method has been chosen *a priori*, it is more efficient to impute at the protein-level."

7. Then, on the basis of 5. and 6., the observed difference in the overall comparison is mainly explained by the performances of the imputation on the 40% or less remaining MCAR values.

8. From 6. and 7., one derives that on these MCAR values, implicit protein-level imputation gives more accurate results.

9. Then combining 8. and 2. leads to the aforementioned conclusion: on MCAR, an implicit and sub-optimal imputation is more efficient than a MNAR-devoted method.

As here, the implicit imputation of the aggregation is equivalent to a mean imputation (which can be seen as a poor MCAR method), it highlights that a bad MCAR method is more efficient on MCAR data than a good MNAR method. While this conclusion may appear trivial, it however stresses that the adequation between the nature of the missing values and the imputation strategy is more important than the theoretical performances (*i.e.* regardless the nature of missing values) of the imputation algorithm.

In addition, a last conclusion can be drawn: as the implicit imputation performed during the aggregation mainly operates on MCAR, so that mainly MNAR remain at protein-level, our results support the idea that the MNAR ratio is generally more important at protein-level than at peptide-level (such as observed in [29, 30] for instance). However, this last conclusion must be cautiously interpreted: indeed, it does not mean that if there are a lot of MNAR, it is better to work at protein-level: to derive such a conclusion, one would need to have mainly red cells in the upper lines of Figures 5 and 6 graphics; yet it only holds for a couple of them (Figures 6(a) and (b)), so that no general conclusion can be drawn.

Of course, if one changes the aggregation method, the comparison between peptide-level and protein-level imputations will lead to slightly different results, and we do not pretend to be exhaustive. However, even if the aggregation strategy is more elaborated than the three aforementioned ones (sum, mean or max), the conclusions are of the same spirit: whatever the aggregation function, it is most likely to have a neutral element that will act as the implicit imputation value, on the basis of which most of the aforementioned conclusions are elaborated.

5 Conclusions

Let us first summarize the conclusions of this work into four points. (1) Imputation should be performed at the peptide-level, since aggregating peptides into proteins beforehand amounts to performing a first implicit and in most of the cases, sub-optimal imputation. (2) In the absence of knowledge about the nature(s) of missing values in a particular quantitative proteomics dataset, it makes sense to rely on a MCAR/MAR imputation method. This is supported by numerous experiments, including ours as well as those from [13], but also by theoretical arguments: by definition, missing values that should be imputed by small intensities can also show up in a MCAR context (so that they can also be imputed to some extent by MCAR-devoted imputation methods), while, on the contrary, a method devoted to left-censored missing value will systematically perform poorly on other types of missing values. (3) However, this conclusion should be moderated by the observation that the superiority of MAR/MCAR-devoted methods only holds on the average and should be contextualized, as cases arise where MNAR-devoted methods perform better than MCAR-devoted ones. Similarly, it appears that choosing a method adapted to the nature of the missing values is more important than choosing a method itself, regardless the nature of missing values. As a consequence, before any imputation, the practitioner should identify the main or most likely nature among the missing values in his/her quantitative dataset, and impute accordingly. (4) Finally, while MNAR are best imputed by specific methods, other missing values are well accounted for by MAR/MCAR-devoted methods. As it is accepted that many types of missing values coexist in most of the quantitative datasets (see for instance [13]), hybrid strategies (based on both MNAR- and MAR/MCAR-devoted methods) should be considered in the future.

These elements shed a new light on the directions that methodological research should follow with regards to missing value imputation in quantitative proteomics. MNAR-devoted methods, that are less numerous and that have been less investigated in the general field of statistics, remain a subject of likely improvements. Concomitantly, important room is left to develop

diagnosis tools, that are capable of categorizing the missing values according to the mechanism that generated them. This diagnosis can operate at different levels: (*i*) at the dataset level, so that the imputation strategy is applied conditionally to the majority nature of missing values in the entire dataset; (*ii*) at the peptide-level, so that all the missing values within a same peptide (in a given group of replicates) are assumed to be of a same nature; (*iii*) at the missing value level, so as to have a most refined categorization of the missing values across the dataset. Finally, once such diagnosis tools are available, it will be possible to elaborate hybrid strategies, that process each group of missing values according to its nature, so as to best preserve the biological relevance of the quantitative datasets and of the biological conclusions.

6 Acknowledgements

This work was supported by the following funding: ANR-2010-GENOM-BTV-002-01 (Chloro-Types), ANR-10-INBS-08 (ProFI project, "Infrastructures Nationales en Biologie et Santé", "Investissements d'Avenir"), EU FP7 program (Prime-XS project, Contract no. 262067), the Prospectom project (Mastodons 2012 CNRS challenge) and the BBSRC Strategic Longer and Larger grant (Award BB/L002817/1).

References

[1] David A. Stead, Norman W. Paton, Paolo Missier, Suzanne M. Embury, Cornelia Hedeler, Binling Jin, Alistair J. P. Brown, and Alun Preece. Information quality in proteomics. *Briefings in Bioinformatics*, 9(2):174–188, 2008.

[2] Daniela Albrecht, Olaf Kniemeyer, Axel A. Brakhage, and Reinhard Guthke. Missing values in gel-based proteomics. *PROTEOMICS*, 10(6):1202–1211, 2010.

[3] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[4] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[5] Yuliya Karpievitch, Alan Dabney, and Richard Smith. Normalization and missing value imputation for label-free lc-ms analysis. *BMC Bioinformatics*, 13(Suppl. 16:S5):1–9, 2012.

[6] Trevor Hastie, Robert Tibshirani, Balasubramanian Narasimhan, and Gilbert Chu. impute: Imputation for microarray data. *R package, version 1.42.0*.

[7] Cosmin Lazar. imputeLCMD: A collection of methods for left-censored missing data imputation. *R package, version 2.0*.

[8] Wolfram Stacklies, Henning Redestig, Matthias Scholz, Dirk Walther, and Joachim Selbig. pcamethods - a bioconductor package providing pca methods for incomplete data. *Bioinformatics*, 23(9):1164–1167, 2007.

[9] Laurent Gatto and Kathryn S Lilley. Msnbase-an r/bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, 28(2):288–289, 2012.

[10] Lei Nie, Gang Wu, Fred J Brockman, and Weiwen Zhang. Integrated analysis of transcriptomic and proteomic data of desulfovibrio vulgaris: zero-inflated poisson regression models to predict abundance of undetected proteins. *Bioinformatics*, 22(13):1641–1647, 2006.

[11] Wandaliz Torres-García, Weiwen Zhang, George C Runger, Roger H Johnson, and Deirdre R Meldrum. Integrative analysis of transcriptomic and proteomic data of desulfovibrio vulgaris: a non-linear model to predict abundance of undetected proteins. *Bioinformatics*, 25(15):1905–1914, 2009.

[12] Wandaliz Torres-Garcia, Steven D Brown, Roger H Johnson, Weiwen Zhang, George C Runger, and Deirdre R Meldrum. Integrative analysis of transcriptomic and proteomic data of shewanella oneidensis: missing value imputation using temporal datasets. *Molecular BioSystems*, 7(4):1093–1104, 2011.

[13] Bobbie-Jo M Webb-Robertson, Holli K Wiberg, Melissa M Matzke, Joseph N Brown, Jing Wang, Jason E McDermott, Richard D Smith, Karin D Rodland, Thomas O Metz, Joel G Pounds, et al. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of proteome research*, 14(5):1993–2001, 2015.

[14] Yuliya Karpievitch, Jeff Stanley, Thomas Taverner, Jianhua Huang, Joshua N. Adkins, Charles Ansong, Fred Heffron, Thomas O. Metz, Wei-Jun Qian, Hyunjin Yoon, Richard D. Smith, and Alan R. Dabney. A statistical framework for protein quantitation in bottom-up ms-based proteomics. *Bioinformatics*, 25(16):2028–2034, 2009.

[15] So Young Ryu, Wei-Jun Qian, David G Camp, Richard D Smith, Ronald G Tompkins, Ronald W Davis, and Wenzhong Xiao. Detecting differential protein expression in large-scale population proteomics. *Bioinformatics*, 30(19):2741–2746, 2014.

[16] Frank Koopmans, L Niels Cornelisse, Tom Heskes, and Tjeerd MH Dijkstra. Empirical bayesian random censoring threshold model improves detection of differentially abundant proteins. *Journal of proteome research*, 13(9):3871–3880, 2014.

[17] Wen Zhang, Yuhong Wei, Vladimir Ignatchenko, Lei Li, Shingo Sakashita, Nhu-An Pham, Paul Taylor, Ming Sound Tsao, Thomas Kislinger, and Michael F. Moran. Proteomic profiles of human lung adeno and squamous cell carcinoma using super-silac and label-free quantification approaches. *PROTEOMICS*, 14(6):795–803, 2014.

[18] Juan A Vizcaíno, Eric W Deutsch, Rui Wang, Attila Csordas, Florian Reisinger, Daniel Ríos, José A Dianes, Zhi Sun, Terry Farrah, Nuno Bandeira, et al. Proteomexchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology*, 32(3):223–226, 2014.

[19] Joseph G Ibrahim, Ming-Hui Chen, Stuart R Lipsitz, and Amy H Herring. Missing-data methods for generalized linear models. *Journal of the American Statistical Association*, 100(469):332–346, 2005.

[20] Joseph L. Schafer and John W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002.

[21] J.L. Schafer. *NORM: Analysis of incomplete multivariate data under a normal model*. University Park, PA: The Methodology Center, The Pennsylvania State University, version 3 edition, 2008.

[22] Jonas S. Almeida, Romesh Stanislaus, Ed Krug, and John M. Arthur. Normalization and analysis of residual variation in two-dimensional gel electrophoresis for quantitative differential proteomics. *PROTEOMICS*, 5(5):1242–1249, 2005.

[23] Sreelatha Meleth, Jessy Deshane, and Helen Kim. The case for well-conducted experiments to validate statistical protocols for 2d gels: different pre-processing = different lists of significant proteins. *BMC Biotechnology*, 5(1):1 – 15, 2005.

[24] Jean-François Chich, Olivier David, Fanny Villers, Brigitte Schaeffer, Didier Lutomski, and Sylvie Huet. Statistics for proteomics: Experimental design and 2-de differential analysis. *Journal of Chromatography B*, 849(1 - 2):261 – 272, 2007.

[25] I. Wasito and B. Mirkin. Nearest neighbour approach in the least-squares imputation algorithms. *JOURNAL OF INFORMATION SCIENCES*, 169:1–25, 2005.

[26] Roderick J. A. Little. Regression with missing x's: A review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992.

[27] Sunghee Oh, Dongwan D. Kang, Guy N. Brock, and George C. Tseng. Biological impact of missing-value imputation on downstream analyses of gene expression profiles. *Bioinformatics*, 27(1):78–86, 2011.

[28] Hua Chen, Chong-Yu Xu, and Shenglian Guo. Comparison and evaluation of multiple gcms, statistical downscaling and hydrological models in the study of climate change impacts on runoff. *Journal of Hydrology*, 434 - 435(0):36 – 45, 2012.

[29] Myriam Ferro, Sabine Brugière, Daniel Salvi, Daphné Seigneurin-Berny, Lucas Moyet, Claire Ramus, Stéphane Miras, Mourad Mellal, Sophie Le Gall, Sylvie Kieffer-Jaquinod, et al. At_chloro, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. *Molecular & Cellular Proteomics*, 9(6):1063–1084, 2010.

[30] Martino Tomizioli, Cosmin Lazar, Sabine Brugière, Thomas Burger, Daniel Salvi, Laurent Gatto, Lucas Moyet, Lisa M Breckels, Anne-Marie Hesse, Kathryn S Lilley, et al. Deciphering thylakoid sub-compartments using a mass spectrometry-based approach. *Molecular & Cellular Proteomics*, 13(8):2147–2167, 2014.

[31] Petra L. Roulhac, James M. Ward, J. Will Thompson, Erik J. Soderblom, Michael Silva, M. Arthur Moseley, and Erich D. Jarvis. Microproteomics: Quantitative proteomic profiling of small numbers of laser-captured cells. *Cold Spring Harbor Protocols*, 2011(2):218–234, 2011.

[32] Christina Ludwig, Manfred Claassen, Alexander Schmidt, and Ruedi Aebersold. Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry. *Molecular and Cellular Proteomics*, 11(3):M111 – 013987, 2012.

[33] Jonas Grossmann, Bernd Roschitzki, Christian Panse, Claudia Fortes, Simon Barkow-Oesterreicher, Dorothea Rutishauser, and Ralph Schlapbach. Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods. *Journal of Proteomics*, 73(9):1740 – 1746, 2010.

[34] Jeffrey C. Silva, Marc V. Gorenstein, Guo-Zhong Li, Johannes P. C. Vissers, and Scott J. Geromanos. Absolute quantification of proteins by lcmse : A virtue of parallel ms acquisition. *Molecular and Cellular Proteomics*, 5(1):144–156, 2006.

List of Figures

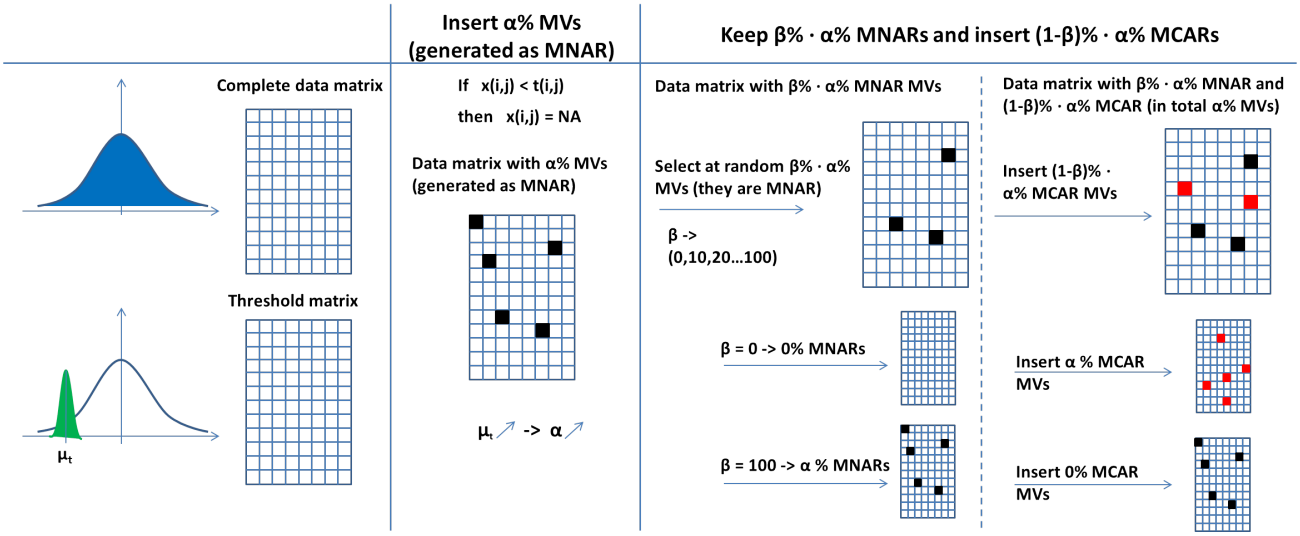
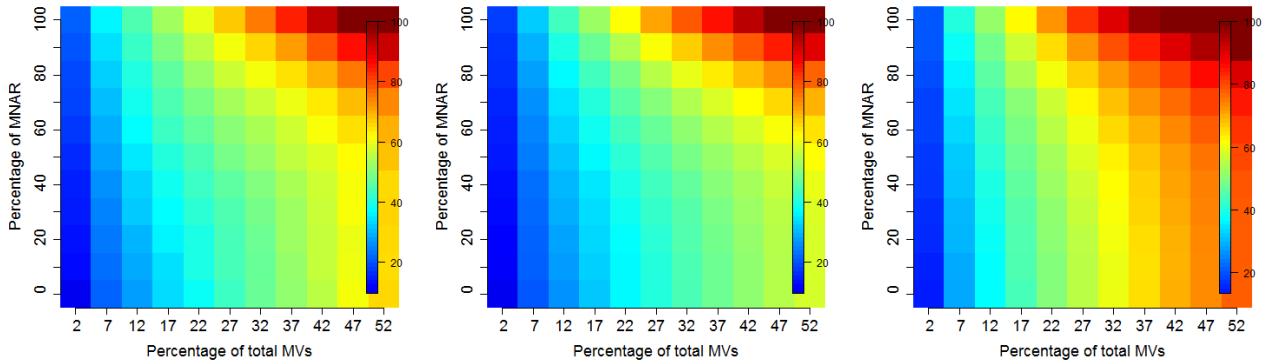
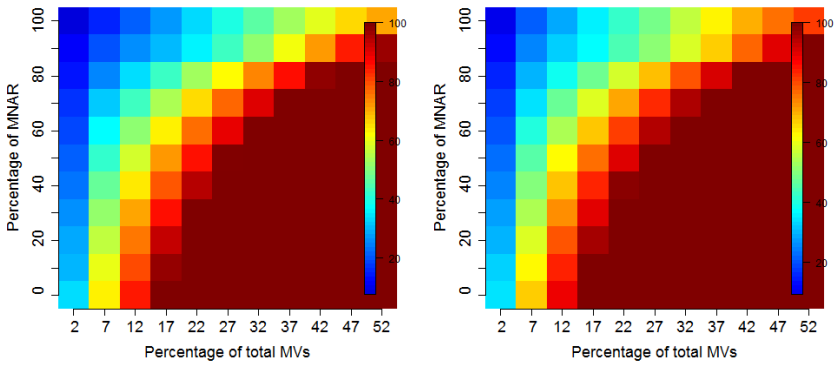


Figure 1: Schematic view upon the strategy used for the missing data generation. This strategy allows to control both for the total proportion of missing values generated, as well as for the proportion of missing values which are MNAR and MCAR.

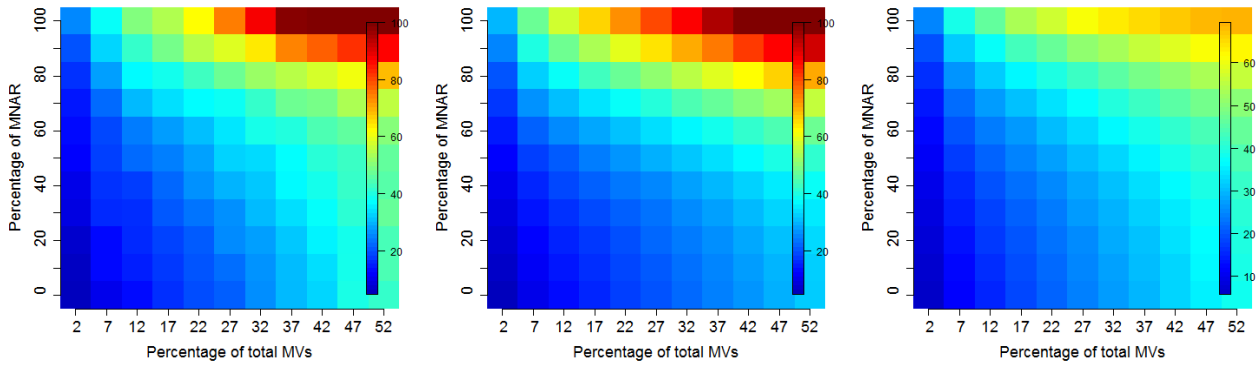


(a) kNN (b) SVDimpute (c) MLE

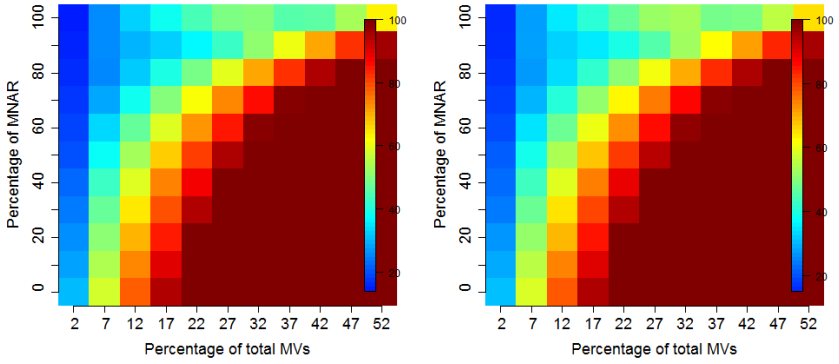


(d) MinDet (e) MinProb

Figure 2: RSR for the simulated quantitative dataset; imputation is performed by considering: kNN (a), SVDimpute (b), MLE (c), MinDet (d) and MinProb (e).

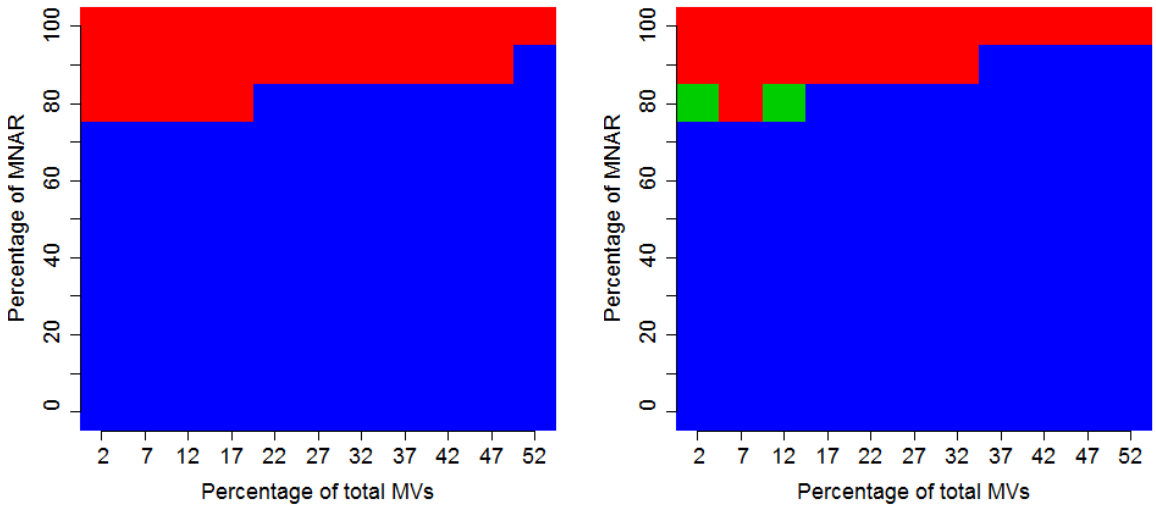


(a) kNN (b) SVDimpute (c) MLE



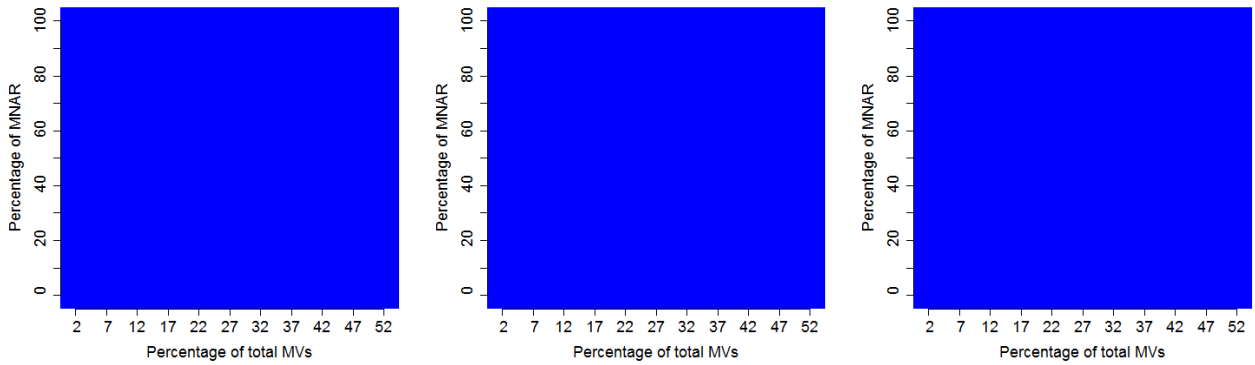
(d) MinDet (e) MinProb

Figure 3: RSR for the real quantitative dataset; imputation is performed by considering: kNN (a), SVDimpute (b), MLE (c), MinDet (d) and MinProb (e).

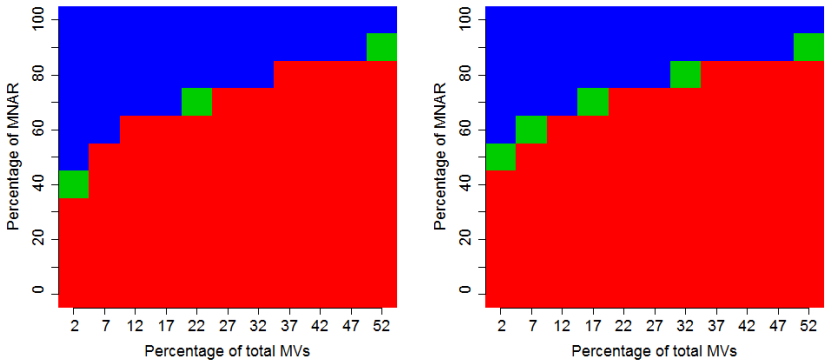


(a) (b)

Figure 4: (a) Comparison of SVDimpute and MinDet on the simulated dataset; (b) Comparison of MLE and MinDet on the real dataset. A red color indicate an outperformance of MinDet, a blue color, an underperformance of MinDet, and a green color, a difference of performance which is not significant with a p-value threshold of 5%.

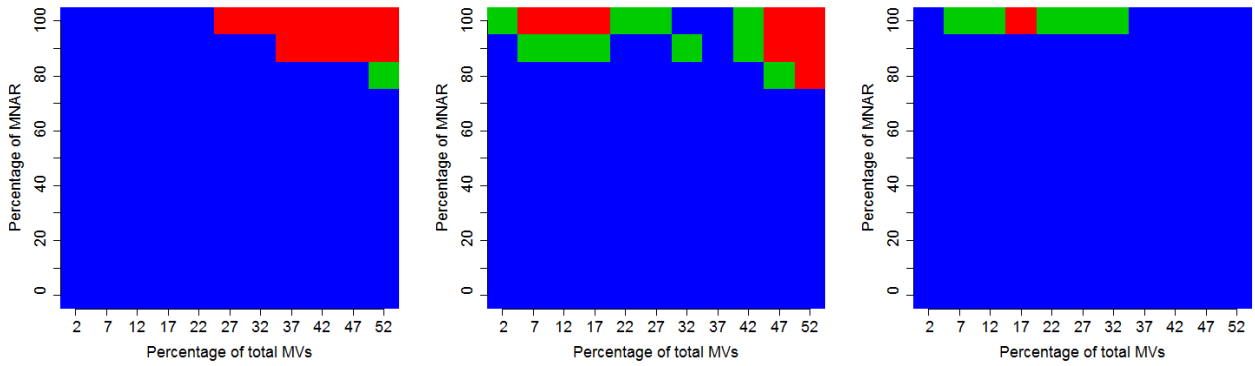


(a) kNN (b) SVDimpute (c) MLE

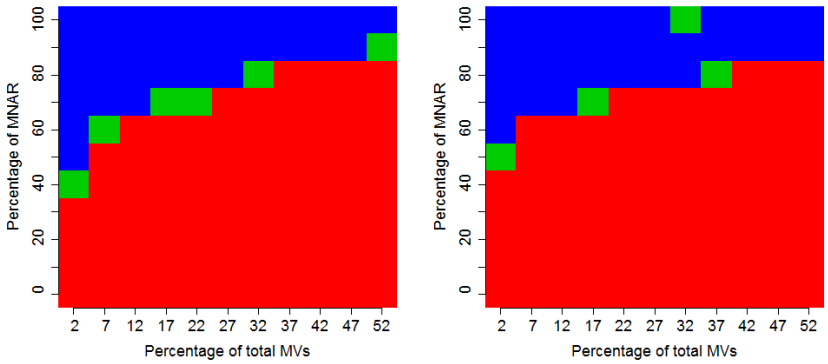


(d) MinDet (e) MinProb

Figure 5: Comparison of peptide-level and protein-level imputations for the simulated quantitative dataset; imputation is performed by considering: kNN (a), SVDimpute (b), MLE (c), MinDet (d) and MinProb (e). Blue indicates peptide imputation superiority, red indicates protein imputation superiority, and green indicates a non-significant result (at 5% threshold).



(a) kNN (b) SVDimpute (c) MLE



(d) MinDet (e) MinProb

Figure 6: Comparison of peptide-level and protein-level imputations for the real quantitative dataset; imputation is performed by considering: kNN (a), SVDimpute (b), MLE (c), MinDet (d) and MinProb (e). Blue indicates peptide imputation superiority, red indicates protein imputation superiority, and green indicates a non-significant result (at 5% threshold).

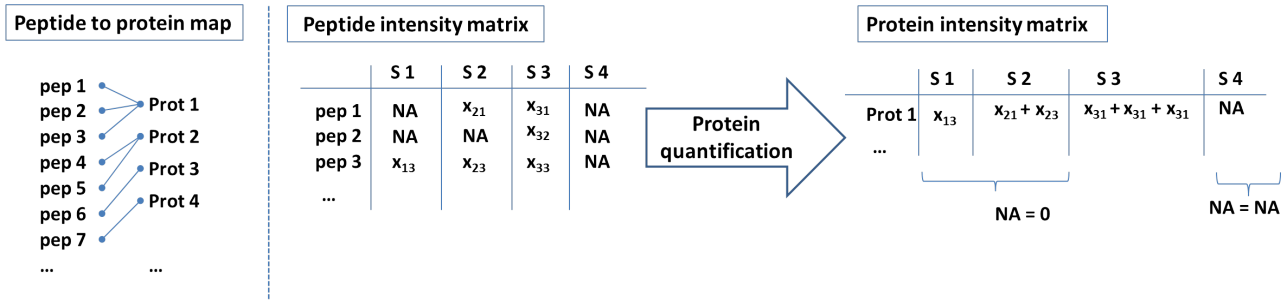


Figure 7: Illustration of implicit missing value imputation during protein quantification from peptide intensity. Here the protein quantification is considered to be performed by summing the signal intensities of all peptides per protein.