基因组学与应用生物学，2021 年，第 40 卷，第 2 期，第 909-915 页

研究报告
Research Report

蛋白质质谱数据缺失值填补方法的比较与评估

郭益浩　李婧 *
上海交通大学生命科学技术学院, 上海, 200240
* 通信作者, jing.li@sjtu.edu.cn

摘　要　基于质谱数据的蛋白质定量分析一直是目前高通量蛋白质组学的重要研究手段。但是基于现有质谱技术的限制，大规模蛋白质定量过程中往往会产生大量的缺失值，这在一定程度上影响了下游分析的准确性。尽管很多缺失值填补方法被不断提出，但是蛋白质组学领域对于不同情况下缺失值填补方法效力的综合评估仍然缺乏。本研究基于真实数据的分布特征，构建模拟数据集，在样本量、效应值以及缺失比例这三个维度上，综合评估了 kNN、SVD、MLE、BPCA、LLS、Min、QRILC、Mean 这 8 种缺失值填补方法的效力。结果显示，填补效力与样本量和效应值呈正相关，也与缺失比例呈负相关。同时，还发现在不同数据集中填补方法的效力有所差异，研究者需要根据数据集特征和自身需求选择适合的填补方法。本研究总结了不同数据集特征下的最优填补方法，供研究者进行参考和使用。

关键词　质谱定量分析; 蛋白质组学; 缺失值; 缺失值填补

Comparison and Evaluation of Imputation Methods for Missing Values in Quantitative Data of Protein Mass Spectrometry

Guo Yihao　Li Jing *
School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, 200240
* Corresponding author, jing.li@sjtu.edu.cn
DOI: 10.13417/j.gab.040.000909

Abstract　Quantitative protein analysis based on mass spectrometry is an important research methodology for high-throughput proteomics. However, due to the limitations of existing mass spectrometry techniques, large-scale quantification process may produce a large number of missing values, which will affect the accuracy of downstream analysis. Although many imputation methods have been proposed, comprehensive evaluation upon those methods in different situations is still lacking in proteomics. Here, based on the characteristics of real data, we constructed different simulation datasets in three dimensions of sample size, effect value, and missing ratio. Then we comprehensively evaluated the imputation effectiveness and accuracy of eight classical methods including kNN, SVD, MLE, BPCA, LLS, Min, QRILC and Mean. The results illustrated that the effectiveness of missing imputation is positively correlated to sample size and effect value, while negatively correlated to missing proportion. We also found that the effectiveness of those methods is varied in different datasets. Researchers need to choose a suitable imputation method according to the characteristics of dataset and their own needs. In this research, we summarize the optimal methods for different characteristics to provide reference for researchers.

Keywords　Mass spectrometry; Proteomics; Missing value; Imputation

基金项目：本研究由国家自然科学基金项目(31871329)和上海市自然科学基金项目(17ZR1413900)共同资助
引用格式：Guo Y.H., and Li J., 2021, Comparison and evaluation of imputation methods for missing values in quantitative data of protein mass spectrometry, Jiyinzuxue yu Yingyong Shengwuxue (Genomics and Applied Biology), 40(2): 909-915. (郭益浩, 李婧, 2021, 蛋白质质谱数据缺失值填补方法的比较与评估, 基因组学与应用生物学, 40(2): 909-915.)

近十几年来，随着蛋白质质谱技术不断发展，日益成熟的蛋白质搜库软件可以较为稳定和可靠地将谱图与已知蛋白质序列进行匹配，使蛋白质与肽段的定性定量分析成为可能。蛋白质质谱技术已然成为了蛋白质定性和定量的重要研究方法。然而，由于技术水平的限制，现有研究方法所获得的定量数据中经常会出现大量的缺失值，尤其是 label-free 定量技术，其定量矩阵缺失比例往往在 30%以上(Albrecht et al., 2010)。而这些缺失值则会对后续的研究分析，尤其是差异表达分析的准确性造成重要影响。因此，蛋白质组学研究的主要挑战之一是适当地处理这些缺失的数据。此外，考虑到许多统计方法都需要完整的数据集，研究中通常会忽略含有缺失值的数据，仅处理能观测到的数据或者使用简单或复杂的数学模型来填补缺失值(Webb-Robertson et al., 2015)。但对于蛋白质组学数据集，往往多于 70%的蛋白质中至少存在一个缺失值，而缺失比例大于 50%的蛋白质也并不罕见(Albrecht et al., 2010)。单纯忽略缺失值将极大地减少数据集的大小和完整性，并限制研究人员推断蛋白质相关信息的能力。

在过去的十几年中，在蛋白质组学领域已发表诸多关于缺失值填补方法及不同填补方法效力评估的报道。Wu 和 Jhou (2017)、Webb-Robertson 等(2015)的研究都全面地比较并讨论了蛋白质组学应用背景下的一些著名填补方法，因此可以从该研究或其中的参考文献中得出诸多结论。已经开发的各种填补方法可分为三类：通过单一数值替换进行填补、基于数据集中的局部结构进行推算填补以及基于数据集中的全局结构进行填补。针对以上三种分类，本研究评估了 kNN (Troyanskaya et al., 2001)、SVD (Troyanskaya et al., 2001)、MLE (Lazar et al., 2016)、BPCA (Stacklies et al., 2007)、LLS (Stacklies et al., 2007)、Min、Mean、QRILC (Wei et al., 2018)等 8 种不同填补方法的效力。其中 Min、Mean 属于单一数值填补法；kNN、LLS 为基于局部结构的方法；MLE、SVD、BPCA、QRILC 为基于全局结构的方法。

迄今为止，对于不同的样本量、效应值(组间丰度差异百分比)以及缺失比例对填补效力的综合评估仍然缺乏。本研究基于真实数据集(Jiang et al., 2019)的数据特征，在样本量和效应值这两个维度上构建完整模拟数据集。然后在此基础上根据不同的缺失比例生成不同的缺失数据集，而后分别使用前述 8 种方法对缺失数据集进行填补，得到每种方法的填补数据集。将这些填补数据集与其对应的完整数据集进行比较，评估在不同数据集特征下各方法缺失填补值的效力，包括数据丰度误差与后续蛋白质差异表达分析的准确性。此外，为了减小随机误差，本研究将每种情况均进行了 100 次重复并取其均值作为最后评估结果。最终讨论并总结了在不同情况下不同缺失填补方法的准确性，为后续蛋白质质谱定量数据中的缺失值处理提供参考。

1 结果与分析

1.1 模拟数据集数据特征选择

为了模拟数据集的真实性和实验结果的普适性，首先观测真实数据集的数据特征，包括其效应值及缺失比例。对于大小为 6 702×198 的真实数据定量矩阵中的每个蛋白质进行了 paired t-test，以检验其组间丰度差异性，并绘制了对应的火山图。其中 *FDR* 为经过 Benjamini 和 Hochberg (1995)校正后的 paired *t*-test *P* value，$\log_2$ (*FC*)为经过对数变换的组间丰度差异。将 *FDR* 小于 0.05，$\log_2$ (*FC*)大于 1 的 1 264个蛋白质作为上调蛋白(红色)，同时将 *FDR* 小于 0.05、$\log_2$ (*FC*)小于 −1 的 232 个蛋白质作为下调蛋白(绿色)，二者数量之和占总体蛋白质数目的 22% (图1A)。此外，还发现存在相当一部分蛋白质，其 *FDR* 小于 0.05，但其 $\log_2$ (*FC*)的绝对值小于 1。

通过计算每个蛋白质在所有样本中的缺失比例，并依据此绘制对应的行缺失比例密度图。根据所示数据发现该真实数据集的整体缺失比例为 21%，50%以上的蛋白质在所有样本中的缺失值比例小于 10%，75%的蛋白质缺失比例小于 40% (图 1B)。

基于真实数据以及相关参考文献(Lazar et al., 2016)和(Langley and Mayr, 2015)中使用的数据集特征，我们最终应用在模拟数据集上的样本量、效应值

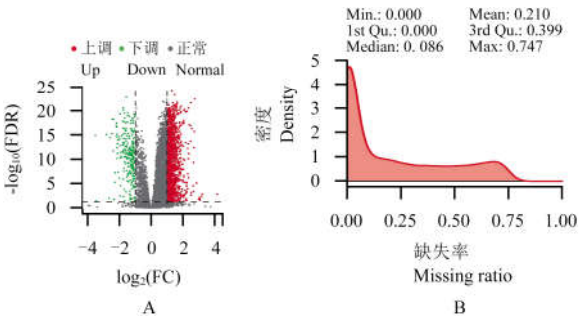

图 1 $\log_2$ (*FC*)分布与行缺失比例分布

注：A: DE 分布; B: 缺失－比率分布(行)

Figure 1 $\log_2$ (*FC*) distribution and row missing ratio distribution

Note: A: DE distribution; B: Missing-ratio distribution (Row)

以及缺失比例的具体取值(表 1)。

1.2 缺失值填补方法的丰度误差比较

以真实数据集为基础，使用 Karpievitch 等(2012)中提出模型的简化版本来构建完整的模拟数据集，其模型如下：

$$Y_{ij}=E_i+G_{ik}+\epsilon_{ij} \quad (1)$$

其中，Y_{ij} 是模拟数据集中第 j 个样本中蛋白质 i 的表达丰度；E_i 是真实数据中蛋白质 i 的平均表达丰度；G_{ik} 是蛋白质 i 在两组间的丰度差异；$\epsilon_{ij}\sim N(0, sd_i)$ 是随机误差项，其中代表了真实数据中蛋白质 i 的丰度方差。以上所用表达丰度均为经过对数转换以及标准化之后的表达丰度。考虑到实际实验中的不同情况，我们在样本量、效应值(组间丰度差异百分比)这两个维度上构建了 36 种完整的模拟数据集。在此基础上，再依据不同的缺失比例生成共 216 种含缺失值的模拟数据集(表 1)。

其中我们还考虑到：基于真实数据集的配对实验设计，模拟数据集中两组别的样本量一致；对于效应值(或组间丰度差异百分比)，可以依据真实数据的情况，从蛋白质总体中随机抽取 20%的蛋白质作为差异表达蛋白加入效应值。

所取用的真实数据为经过 $\log_2$ 对数变换的定量矩阵，所以添加效应值时应为：

$$\log_2(\text{value}\times(1+\text{effectsize}))=\log_2(\text{value})=\log_2(1+\text{effectsize}) \quad (2)$$

每种数据特征重复生成 100 个模拟数据集，结果计算时则通过取 100 次重复的平均值来减小随机误差。

在对含缺失值的模拟数据集进行填补后，本研究还比较了每个填补模拟数据集和其对应的完整模拟数据集之间的丰度差异来计算填补方法的标准均方根误差(normalized root mean square error, NRMSE)，从而评估其丰度误差。在不同的样本量、效应值与缺失比例组合下，8 种缺失值填补方法中 NRMSE 的最小值，即最优填补方法的丰度误差(图 2)。

表 1 模拟数据集数据特征选择范围

Table 1 Construction of simulated data set range

维度 Dimension	取值范围 Ranges
样本量 Sample size	3+3, 5+5, 10+10, 30+30, 50+50, 100+100
效应值(%) Effect size (%)	50, 80, 100, 150, 200, 500
缺失比例(%) Missing ratio (%)	10, 20, 30, 40, 50, 60

注: 总样本量等于正常组样本量+癌症组样本量

Note: The total sample size is equal to the normal group sample size+cancer group sample size

首先，从整体上可以看出，在样本量和缺失比例不变的情况下，每种情况的最优 NRMSE 随效应值的变化波动并不明显，因此推测最优填补方法的 NRMSE 与效应值并不相关。还发现在(C)~(E)中NRMSE 在红实线附近存在明显下降。我们依据此，即梯度，将所有数据结果分为小样本量与大样本量两组分。

小样本组对应的情况，基于 kNN 的填补方法均是最优的填补方法。结果发现，在该组中(图 2A; 图2B)，随着缺失比例的提高，NRMSE 也会随之提高，同时样本量的增加也会导致 NRMSE 的升高。在大样本组中的所有情况下 NRMSE 均基本相似，因此推测在大样本量的情况下，NRMSE 与样本量、较低效应值以及缺失比例关系均不大。对于大样本组我们推荐使用基于 BPCA 的填补方法进行缺失值填补。

最后，考虑到两个分组之间 NRMSE 差异较大，所以推荐使用样本量大于 30 的实验设计，这样能够有效地减小由缺失值填补带来的丰度误差。

1.3 缺失值填补方法的差异表达准确性比较

本研究中，为了探究缺失值填补方法在下游蛋白质差异表达分析准确性上的效力，通过分别对填补数据集以及其对应的完整模拟数据集进行了差异表达分析，并计算同一情况下 8 种缺失值填补方法差异表达分析的生物标志物一致性系数(biomarker list concordance index, BLCI)。我们挑选展示了每种情况下所有缺失值填补方法的最大 BLCI 值及其对应的最优方法(图 3)。

结果显示(图 3A~ 图 3E)，最优 BLCI 和样本量、效应值这两个维度成正相关，和缺失比例大小成负相关。尽管这是预期的结果，但该结果从一定程度上说明模拟方法的有效性。然而过小的样本量将严重影响差异分析的准确性，仅有在缺失比例很低的情况下其最优 BLCI 才能达到 80% (图 3A; 图 3B)。与此同时，BLCI 的增加幅度是存在边际效应的，随着样本量或效应值的大幅度增加，虽然最优 BLCI 仍是增加，但提升幅度却是在逐渐减小，这种现象在 10%~30%的缺失比例上尤为明显(图 3D~ 图 3F)。同时还发现对于不同的数据集特征，均有最适合的缺失值填补方法，但对于效应值大于 100% ($\log_2(FC)>1$)，缺

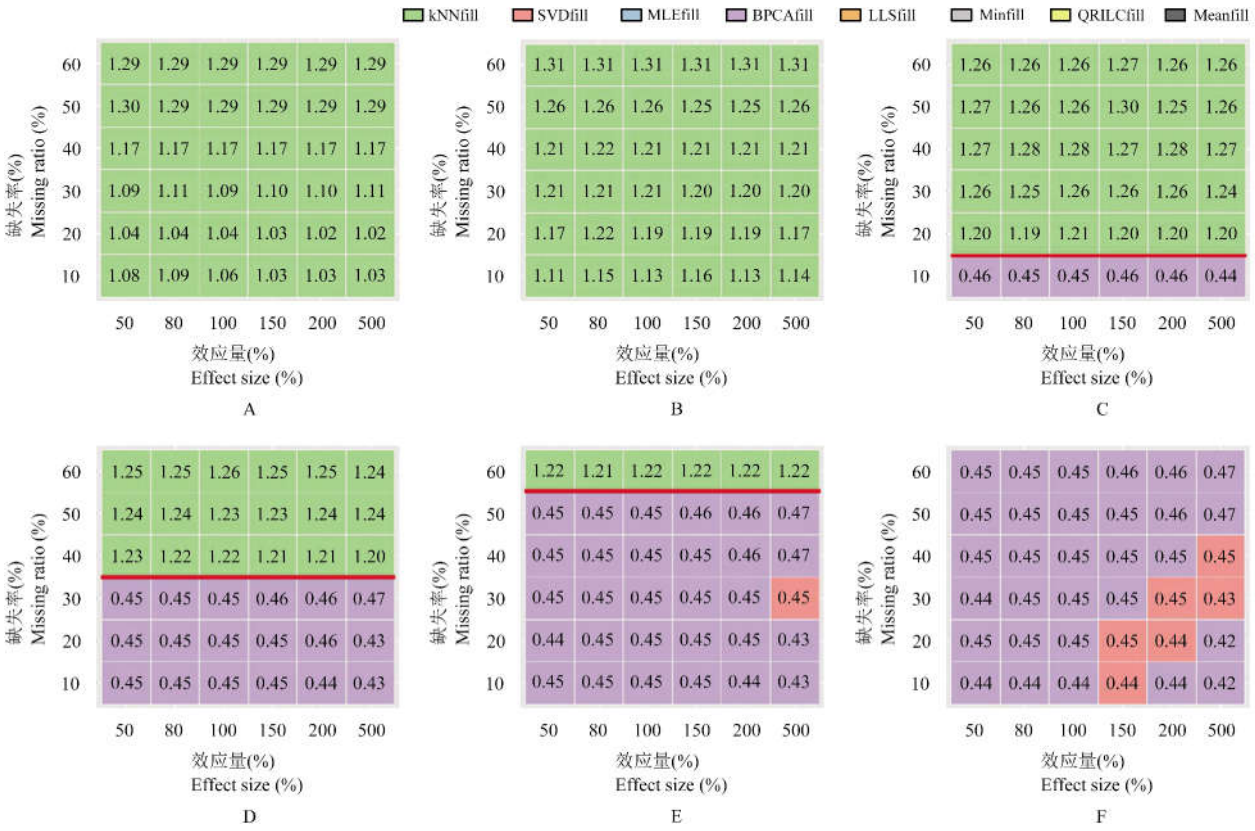

图 2 最优 NRMSE 对应的缺失值填补方法

注: A: 3−3 样品 NRMSE; B: 5−5 样品 NRMSE; C: 10−10 样品 NRMSE; D: 30−30 样品 NRMSE; E: 50−50 样品 NRMSE; F: 100−100 样品 NRMSE

Figure 2 Missing value imputation method corresponding to the optimal NRMSE

Note: A: 3−3 Samples NRMSE; B: 5−5 Samples NRMSE; C: 10−10 Samples NRMSE; D: 30−30 Samples NRMSE; E: 50−50 Samples NRMSE; F: 100−100 Samples NRMSE

失比例小于 30%时，kNN 填补在大样本层面上表现更优异，LLS 填补在小样本层面上表现较为优异。因此，为了能够使差异表达分析尽可能的准确，我们推荐尽量控制定量矩阵的缺失比例小于等于 20%，选取效应值大于 100%，即差异丰度大于 1 倍的蛋白质，此时若样本量大于 10+10 即可保证相对较好的填补效力。

结合前一节以 NRMSE 为评判标准的结果来看，我们发现不同的标准之间，同样情况下的最优填补方法也不一定相同。但二者总体趋势基本一致，对于样本量大、效应值高、缺失比例低的情况填补效果就会更好。对于不同的情况，以及不同的评估标准，我们的结论在处理质谱定量数据的缺失值上有一定的参考价值。

2 讨论

本研究综合评估了 kNN、SVD、MLE、BPCA、LLS、Min、QRILC、Mean 等 8 种不同缺失填补方法在不同数据集特征下的填补效力。发现缺失值填补效力会随着样本量的增加而显著提升，与此同时，其与效应值成正相关，与缺失比例成负相关。结合本研究结论，不推荐使用 3~5 个样本量的实验设计，过少的样本中仅 10%的缺失比例都会显著影响差异表达准确率。若要使填补数据丰度更接近真实数据，则在大样本情况下使用 BPCA 填补，小样本情况下使用 KNN 填补；若为了后续差异表达分析更加准确，则推荐删除缺失比例过高的蛋白质使整体矩阵的缺失比例控制在 30%以下，此时对于大样本使用 kNN 填补，对小样本使用 LLS 填补。但对于所有情况，我们发现从总体上看并没有一种缺失填补方法在所有情况下均表现得最好，我们的研究提供了每种情况最适合的填补方法(图 2; 图 3)。

迄今为止，针对蛋白质组学领域研究者们已发表诸多关于缺失值填补方法及多种方法效力评估的文章，但对于样本量、效应值以及缺失比例这三个层面上的综合评估仍然缺乏。本研究通过建立模拟定

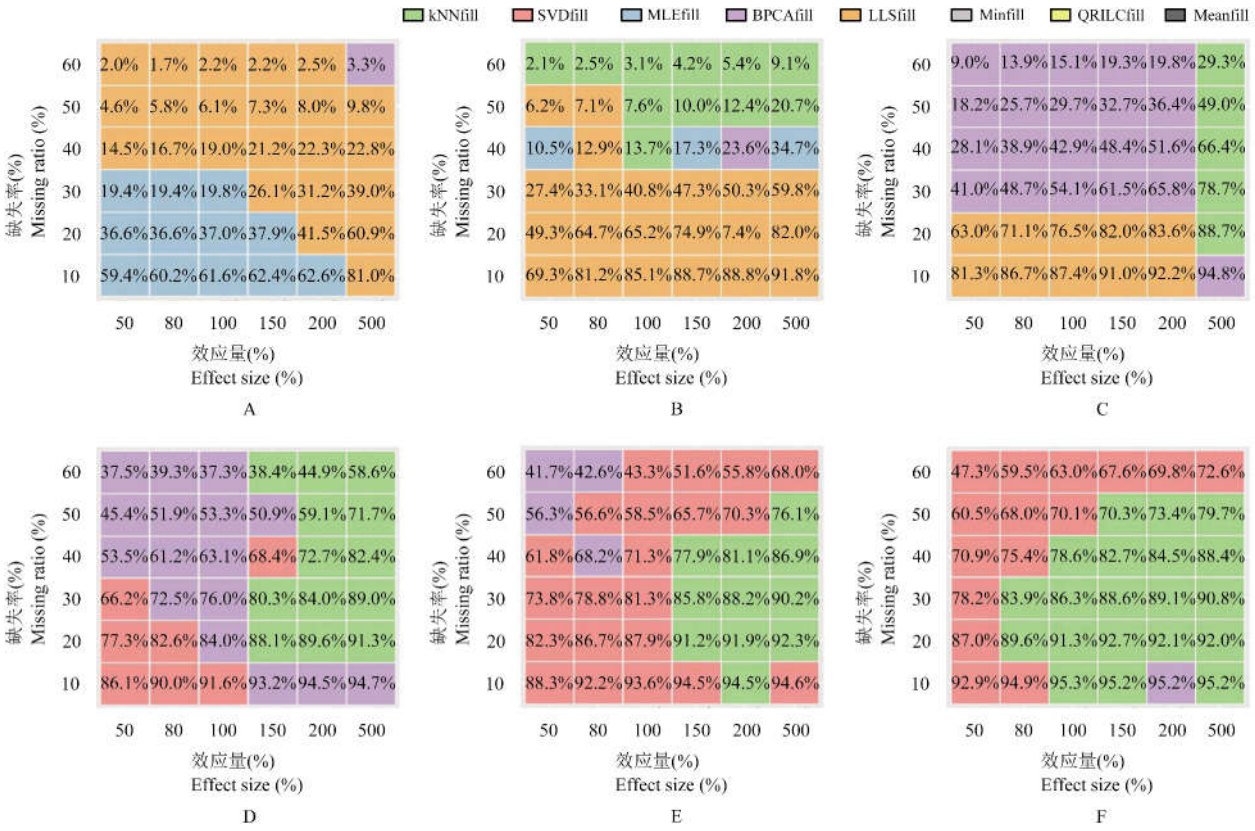

图 3 最优 BLCI 对应的缺失值填补方法

注: A: 3−3 样品 BLCI; B: 5−5 样品 BLCI; C: 10−10 样品 BLCI; D: 30−30 样品 BLCI; E: 50−50 样品 BLCI; F: 100−100 样品 BLCI

Figure 3 Missing value imputation method corresponding to the optimal BLCI

Note: A: 3−3 Samples BLCI; B: 5−5 Samples BLCI; C: 10−10 Samples BLCI; D: 30−30 Samples BLCI; E: 50−50 Samples BLCI; F: 100−100 Samples BLCI

量矩阵综合地评估了 8 种缺失值填补方法在不同数据集特征下的填补效力，为质谱定量数据的缺失值填补提供了重要的参考，不仅可以帮助研究者们进行数据预处理时提供参考，也可以为其他领域的缺失值填补效力评估提供新思路。此外，对于实际研究中的其他没有覆盖到的情况，比如 unpaired 实验设计、不同实验平台数据差异等方面，仍需进一步探究。

3 材料与方法

3.1 数据集来源及预处理

为了更接近真实的构建模拟数据集，选择了一个基于 Label-free 质谱分析的真实公共数据集作为参照标准。该数据集是基于对 110 例早期肝细胞癌症患者的癌症与癌旁组织的研究所得(Jiang et al., 2019)。通过 MaxQuant (1.5.3.30 版本)将原始质谱结果与人类的 Uniprot 数据库(2014 年 9 月发布; http://www.uniprot.ortg)进行蛋白质鉴定。完整的 MaxQuant 结果可从 PRIDE 数据库(www.ebi.ac.uk/pride/archive)中下载得到，其标识符为 PXD006512。

对于搜库后的定量矩阵，我们依据该研究中的方法进行了质量控制、分位数标准化、对数转换以及过滤掉缺失比例大于 75%的蛋白质等预处理步骤。最后我们得到了大小为 6 702×198 的定量矩阵(蛋白质×样本)，用于后续的模拟数据集构建。

3.2 缺失值填补方法

在过去的十几年中，已经开发了诸多应用于质谱数据缺失值的填补方法，但这些算法通常可分为三类：(1)单一数值替换的填补方法；(2)基于数据集局部结构的填补方法；(3)基于数据集整体结构的填补方法(Webb-Robertson et al., 2015)。本研究评估了 8 种缺失值填补方法的效力，所有填补方法均由 R 语言来实现，每种方法对应的描述内容如下：

kNNfill (k 最近邻填补)：通过 k 个最相似蛋白质的加权平均丰度来填补缺失值，基于 imputeLCMD 包(v2.0)实现 (Lazar et al., 2016)。

SVDfill (奇异值分解填补)：使用矩阵的 k 个秩

不断迭代到收敛，进而完成缺失值填补，基于 imputeLCMD 包(v2.0)实现。

MLEfill (最大似然估计填补)：利用最大似然估计原理得出定量数据所服从的未知参数估算量，再完成缺失值填补，基于 imputeLCMD 包(v2.0)实现。

QRILC (左删失数据的分位数回归填补)：通过随机抽取分位数回归估计的截断分布来进行缺失值填补，基于 imputeLCMD 包(2.0 版本)实现。

Minfill (最小值填补)：使用每个列的最小值来对该列的所有缺失值进行填补。

Meanfill (平均值填补)：使用每个蛋白质中非缺失元素的平均值来进行填补。

LLSfill (局部最小二乘填补)：基于 k 个最近蛋白质的多元最小二乘估计填补缺失值，使用 pcaMethods 包(1.72.0 版)实现(Stacklies et al., 2007)。

BPCAfill (基于贝叶斯的主成分分析填补)：利用变分贝叶斯算法来估算模型参数的后验分布再完成缺失值的填补，使用 pcaMethods 包(1.72.0 版)实现。

3.3 评估标准

本研究使用的差异表达方法均为 paired t-test，对其 P value 进行 Benjamini-Hochberg 校正后为 FDR，我们将 FDR<0.05 的蛋白质视为存在显著差异的蛋白质。为了更全面地评估各个缺失值填补方法的效力，从丰度误差和差异表达准确度这两个层面进行了探究。本研究分别使用了完全均方根误差(NRMSE) (Oba et al., 2003)和生物标志物一致性指数(BLCI) (Oh et al., 2011)作为评估标准，其公式内容如下：

$$NRMSE=\sqrt{\frac{mean[(v_{\text{imputed}}-v_{\text{raw}})^2]}{variance[v_{\text{raw}}]}} \quad (3)$$

$$BLCI=TPR+TNR-1 \quad (4)$$

其中，v_{imputed} 为缺失值的填补估算值，v_{raw} 为完整模拟矩阵中对应位置的模拟取值；TPR 为缺失填补方法差异表达分析的真阳性率，TNR 为对应的真阴性率，二者通过比对完整矩阵与填补矩阵的差异表达分析结果所得。

作者贡献

郭益浩是本研究的实验设计者和研究的执行者，完成了方法评估和论文初稿的写作；李婧是项目的构思者及负责人，指导研究设计、数据分析、论文写作与修改。两位作者都阅读并同意最终的文本。

致谢

本研究由国家自然科学基金项目 (31871329)和上海市自然科学基金项目(17ZR1413900)共同资助。

参考文献

Albrecht D., Kniemeyer O., Brakhage A.A., and Guthke R., 2010, Missing values in gel-based proteomics, Proteomics, 10(6): 1202-1211.

Benjamini Y., and Hochberg Y., 1995, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. Roy. Statist. Soc., 57(1): 289-300.

Jiang Y., Sun A.H., Zhao Y., Ying W.T., Sun H.C., Yang X.R., Xing B.C., Sun W., Ren L.L., Hu B., Li C.Y., Zhang L., Qin G.R., Zhang M.H., Chen N., Zhang M.L., Huang Y., Zhou J.N., Zhao Y., Liu M.W., Zhu X.D., Qiu Y., Sun Y.J., Huang C., Yan M., Wang M.C., Liu W., Tian F., Xu H.L., Zhou J., Wu Z.Y., Shi T.L., Zhu W.M., Qin J., Xie L., Fan J., Qian X.H., and He F.C., 2019, Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma, Nature, 567(7747): 257-261.

Karpievitch Y.V., Dabney A.R., and Smith R.D., 2012, Normalization and missing value imputation for label-free LC-MS analysis, BMC Bioinformatics, 13(16): 5.

Langley S.R., and Mayr M., 2015, Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics, J. Proteomics, 129: 83-92.

Lazar C., Gatto L., Ferro M., Bruley C., and Burger T., 2016, Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies, J. Proteome Res., 15(4): 1116-1125.

Oba S., Sato M.A., Takemasa I., Monden M., Matsubara K., and Ishii S., 2003, A Bayesian missing value estimation method for gene expression profile data, Bioinformatics, 19(16): 2088-2096.

Oh S., Kang D.D., Brock G.N., and Tseng G.C., 2011, Biological impact of missing-value imputation on downstream analyses of gene expression profiles, Bioinformatics, 27(1): 78-86.

Stacklies W., Redestig H., Scholz M., Walther D., and Selbig J., 2007, Pcamethods-a bioconductor package providing PCA methods for incomplete data, Bioinformatics, 23(9): 1164-1167.

Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., and Altman R.B., 2001, Missing value estimation methods for DNA microarrays, Bioinformatics, 17(6): 520-525.

Webb-Robertson B.J.M., Wiberg H.K., Matzke M.M.,. Brown J.

N., Wang J., McDermott J.E., Smith R.D., Rodland K.D., Metz T.O., Pounds J.G., and Waters K.M., 2015, Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics, J. Proteome Res., 14(5): 1993-2001.

Wei R.M., Wang J.Y., Su M.M., Jia E., Chen S.Q., Chen T.L., and Ni Y., 2018, Missing value imputation approach for mass spectrometry-based metabolomics data, Sci. Rep., 8(1): 663.

Wu W.S., and Jhou M.J., 2017, MVIAeval: a web tool for comprehensively evaluating the performance of a new missing value imputation algorithm, BME Bioinformatics, 18: 31.