Benchmarking statistical methods for mass spectrometry-based proteomics

Tushita Gupta with Devon Kohler(Ph.D. Candidate)
Advisor: Professor Olga Vitek

Northeastern
Khoury College of
Computer Sciences

Introduction

Our research compares statistical methods for analyzing mass spectrometry proteomics data to help biologists choose the best method for their specific scenario. This is important due to the various processing tools and biological variations that can affect the results, making it challenging to choose an appropriate method. By providing insights into the comparative performance of different methods, we aim to assist biologists in making informed decisions in their data analysis.

Data Overview

The datasets used in our research are summarized in Table 1. Two controlled mixtures and one biological mixture were employed representing the DDA data acquisition strategy. To compare the impact of preprocessing tools on the data, different preprocessing tools were used to generate multiple datasets from a single raw dataset.

Table 1: Datasets used in the research

Dataset	Comparison	Number of Conditions	Number of Biological Replicates	Number of Technical Replicates	Data Processing Tool
Dataset1-DDA: ControlledMix (Controlled)	Group	5	1	3	Skyline MaxQuant Progenesis P.D.
Dataset2-DDA:Choi2017 (Controlled)	Group	4	1	3	Skyline MaxQuant Progenesis
Dataset3-DDA:Meierhofer (Biological)	Paired	2	6	2	MaxQuant

Table 2: Experimental design of the data

Whole plot

Subplot	Condition$_1$						···	Condition$_k$						···
	Subject$_1$			Subject$_2$			··· Subject$_L$	Subject$_{L+1}$			Subject$_{2L}$			··· Subject$_N$
	Run$_1$	Run$_2$	Run$_3$	Run$_4$	Run$_5$	Run$_6$	··· Run$_K$	Run$_{K+1}$	Run$_{K+2}$	Run$_{K+3}$	Run$_{2K}$	···	Run$_{(N-1)K+1}$	Run$_{(N-1)K+2}$	Run$_{(N-1)K+3}$	Run$_{NK}$
Feature$_1$	y	y	y	y	y	y	··· y	y	y	y	···	y	y	y	y	y	y	··· y	y	y
Feature$_2$	y	y	y	y	y	**NA**	··· y	y	y	···	y	y	y	y	y	y	··· y	y	y
···	···	···	···	···	···	···	··· ···	···	···	···	···	···	···	···	···	···	···	··· ···	···	···
Feature$_M$	y	**Cens**	y	**Cens**	**Cens**	y	··· **Cens**	y	y	···	**NA**	y	y	y	y	y	··· y	**Cens**	y

Statistical Methods

- The research compared the performance of eight different statistical methods, namely MSstats, MSqRob2, DEqMS, pmartR, DEP, proDA, prolfqua, and limma.
- The statistical methods were evaluated using controlled mixtures and biological experiments to assess their performance in different scenarios.

Figure 1: Methodology workflow

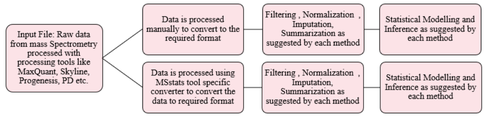


Table 3: Comparative analysis of functionality and workflow of the statistical methods

	MSstats	PmartR	DEP	ProDA	DeqMS	MSqRob2	Prolfqua	Limma
Input File	Two options used: 1. Native preprocessing 2. MSstats converter preprocessing							
Filtering	Based on shared peptides, decoy proteins	Based on proteins, statistical fitting	Based on missing values	None	None	Based on number of peptides	Based on intensity, number of peptides	None
Normalization	Equalize medians	Median	Variance stabilizing transformation	Median	None	Median	Robust z-score	None
Imputation	AFT model	None	MinProb	None	None	None	None	None
Summarization	TMP	Roll up	Already summarized experiment object	Matrix with column per sample and row per protein	None	Aggregate peptides for each protein	None	None
Statistical model	Linear mixed effect model	Anova model	Protein-wise linear models and empirical bayes statistics using limma	Linear probabilistic dropout model	Limma workflow and a function to moderate variance based on feature counts	Linear mixed effects model with empirical bayes moderation	Linear model	Linear model fit with bayes moderation

Results

1 Processing tool affects the performance of statistical methods

Figure 2: Distribution of missing values for dataset1 : controlled mixture

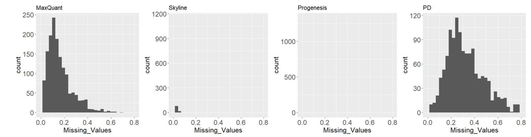
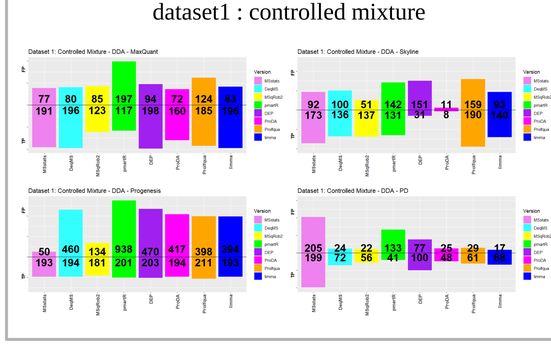
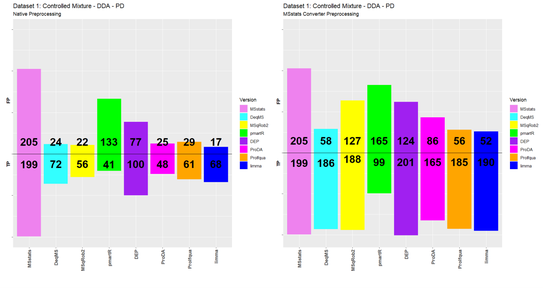


Figure 3: Statistical results for each method for dataset1 : controlled mixture



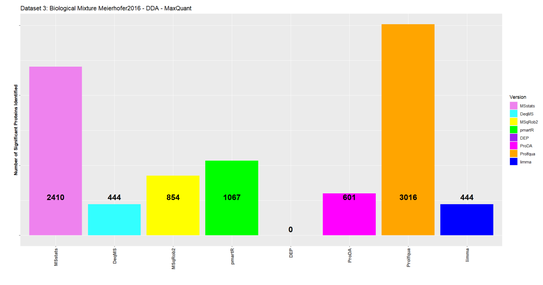
2 Upstream processing affects the performance of statistical methods

Figure 4: Difference in statistical results based on different preprocessing



3 Biological variation affects the performance of statistical methods

Figure 5: Statistical results for each method for dataset 3: biological mixture



Conclusions

- Different processing tools generate different missing values distributions for the same raw data, resulting in different statistical results. This highlights the impact of missing values on statistical methods.
- Upstream processing has a significant influence on statistical inference.
- The presence of biological variation also affects the performance of statistical methods.

Future Directions

- Generate guidelines to benchmark statistical methods for mass spectrometry proteomics data analysis.
- Compare the performance of statistical methods on real and simulated datasets.
- Gain insights into the comparative performance of different methods and provide guidance for biologists in their data analysis.