Unlocking the discriminative potential of MALDI-TOF profiling data with AI-enabled software for biomarker discovery, classification and typing

Luis Mancera[1], **Jesús Jimenez**[1], **Manuel J. Arroyo**[1], **Arndt Asperger**[2]
[1] Clover Bioanalytical Software, S.L. Av. del Conocimiento 41, 18016 Granada, Spain
[2] Bruker Daltonics GmbH & Co. KG, Bremen, Germany

Introduction

Because of its fast analysis speed, high sample throughput, and access to a wide analyte space, MALDI-TOF MS is an ideal tool for rapid and simple profiling analysis allowing for biomarker discovery, classification and typing in a broad range of applications.

We describe here the use of an AI-enabled, workflow-oriented software platform for efficient analysis of MALDI-TOF profiling data in various application examples:

i) Identification of biomarkers in human blood serum protein profiles

ii) Classification (i.e., Hierarchical Clustering / PCA) of research-grade biosimilars based on their N-glycan profiles

iii) Verification of closely related fish species based on protein profiles from tissue extracts applying a supervised machine-learning based classifier model (SVM)

Methods

MALDI-TOF MS data were acquired on neofleX™, autofleX® and microfleX® series instruments (Bruker, Bremen, Germany), respectively. Protein profiles were acquired in positive linear mode, N-glycan profiles in positive reflector mode.

Spectra pre-processing (i.e., smoothing, baseline correction, m/z alignment), m/z feature table generation and statistical data analysis were performed using Clover MSDA software (Clover Bioanalytical Software S.L., Granada, Spain), a cloud-based application accessible via web browser interface. Raw spectra were uploaded in Bruker proprietary format and were organized in a customizable study and experiment structure.

Clover MSDA software offers three tailored analysis workflows:

Biomarker Analysis: Workflow aiming for detection of discriminative m/z features separating samples from known categories. Uses univariate statistical methods, such as T-test, Mann-Whitney U test, and ROC.

Classification: Workflow enabling classification of MALDI-TOF spectra employing various unsupervised and supervised methods (multivariate statistics; clustering; ML based classifiers), such as Principal Component Analysis (PCA), Hierarchical Clustering, K-Means Clustering, Partial Least Squares Discriminant Analysis (PLS-DA), Support Vector Machine (SVM), Random Forest (RF), K Nearest Neighbors (KNN), and Light Gradient-Boosting Machine (LightGBM).

Results from PLS-DA, SVM, RF, KNN and LightGBM can be saved as prediction models for category assignment of samples of unknown origin (*Identification* workflow).

Identification: Workflow intended for category assignment of spectra from unknown samples using one or multiple prediction models generated in the *Classification* workflow

Reference

G.Huffman, L.Mancera, A.Asperger, Bruker Technical Note TN-63, www.bruker.com

Results

Workflow example I - Biomarker Analysis

Experiment: Identification of most discriminative m/z features in MALDI-TOF protein profiles to differentiate between human blood serum samples spiked with artificial peptides and non-spiked controls

MALDI-TOF dataset: 20 spectra per sample category (i.e., spiked, non-spiked), m/z range 1000-10,000, average-mass m/z features

Mass	p-value	q-value (< 0.05)
1297.6	1.66 e-9	4.72 e-8
1348.9	1.46 e-10	8.29 e-9
1621.0	3.63 e-11	3.40 e-9
1637.3	0.0000855	0.0019347
1760.3	2.84 e-10	1.21 e-8
2094.5	3.98 e-11	3.40 e-9
2105.6	0.0024573	0.0466880
2109.5	0.0000905	0.0019347
3148.8	4.19 e-10	1.43 e-8

Peak	AUC (>= 0.8)	Appearance	Positive Category
1297.6	1.000	20 out of 40	20 out of 20
1348.9	1.000	20 out of 40	20 out of 20
1621.0	1.000	21 out of 40	20 out of 20
1637.3	0.800	12 out of 40	12 out of 20
1760.3	1.000	20 out of 40	20 out of 20
2094.5	1.000	21 out of 40	20 out of 20
2109.5	0.800	12 out of 40	12 out of 20
2466.6	0.905	24 out of 40	20 out of 20
3148.8	1.000	20 out of 40	20 out of 20

Figure 1: **Result of biomarker analysis spiked vs. non-spiked human blood serum**
Top left: Butterfly plot of representative MALDI-TOF spectra from non-spiked and spiked human blood serum. Marked peaks represent spiked peptides. **Top right:** Outlier detection result (PCA reconstruction and spectra correlation plots). **Center left:** Peak heatmap. **Center right:** Pearson correlation heatmap. **Bottom left:** T-test result table listing m/z features with q < 0.05. **Bottom right:** ROC result table reporting m/z features with AUC ≥ 0.8.
All m/z features revealed by T-test and ROC analysis as discriminative between sample categories represent the spiked peptides and their oxidized versions.

Workflow example II - Classification (here: Hierarchical Clustering/PCA)

Experiment: Classification of 4 nivolumab research-grade biosimilar samples based on their MALDI-TOF N-glycan profile spectra

MALDI-TOF dataset: 30 spectra per sample, m/z range 1000-3500, monoisotopic m/z features

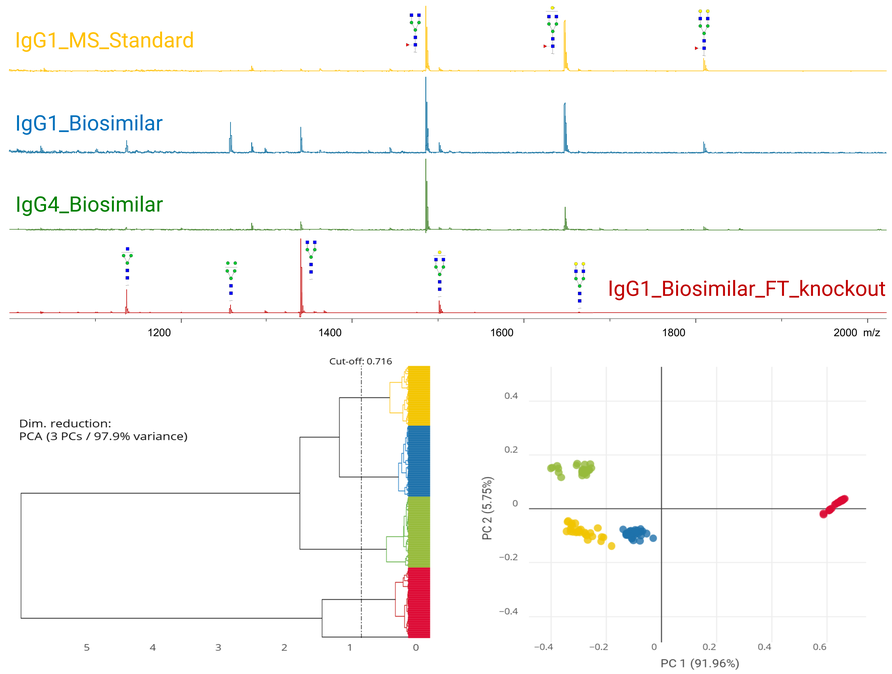

Figure 2: **Classification result obtained for nivolumab research-grade biosimilars**
Top: MALDI-TOF N-glycan spectra from nivolumab research-grade biosimilars. **Bottom left:** Hierarchical clustering result (dendrogram). **Bottom right:** PCA result (score plot).
Hierarchical clustering and PCA successfully classified the spectra by sample origin. Furthermore, both methods confirmed the fucosyltransferase (FT) knockout IgG1 biosimilar being particularly different in its N-glycan profile from the other three biosimilars, which is explained by the lack of core-fucosylated N-glycans in the FT knockout IgG1. Additionally, PCA separated the biosimilar samples by IgG subtype (IgG1 vs. IgG4 Fc regions) along PC2.

Workflow example III – Identification of unknown samples using a trained & validated SVM prediction model

Experiment: Training and validation of an SVM prediction model for verification of closely related *Salmonidae* family fish species based on MALDI-TOF protein profiles generated from fish tissue extracts.

MALDI-TOF dataset: Training: 15-16 spectra per fish species, Validation: 6-8 spectra per fish species, m/z range 2000-30,000, average-mass m/z features

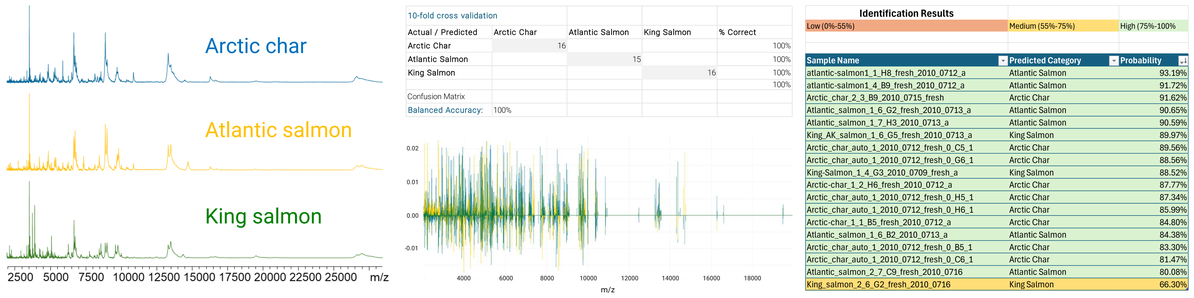

Figure 3: **Verification of *Salmonidae* fish species applying a SVM prediction model**
Left: MALDI-TOF protein profiles of *Salmonidae* fish species. **Center top:** Result from 10-fold cross validation of the SVM model. **Center bottom:** SVM loadings used for model generation. **Right:** Identification results obtained from the SVM model for 18 "unknown" samples. 17 spectra were correctly verified with high probability (>75%), one with medium probability (55%-75%).

Conclusions

- Clover MSDA software provides a powerful statistical toolbox to efficiently leverage the discriminative power of MALDI-TOF MS profiling data.
- Clover MSDA provides access to a wide range of AI-enabled data analysis methods via an intuitive, workflow-oriented web interface. It features user management, project- and experiment-driven data management, as well as detailed reporting of results.
- Clover MSDA supports tailored data analysis workflows allowing for time-efficient analysis of large cohorts of MALDI-TOF spectra with minimal effort enabling simple and fast biomarker discovery, classification and typing across a broad range of life science and further applications.

MALDI-TOF; Classification & Typing

COI Disclosure: L.M., J.J. and M.A. are employees of Clover Bioanalytical Software. Clover develops and sells software used in this study. A.A. is employee of Bruker Corporation. Bruker manufactures and sells instrumentation and software used in this study.