N- and C-terminal Protein Sequencing
Using in-source-decay (ISD) MALDI MS

By Ejvind Mortz, Thanh Ha Nguyen, and Janne Crawford. Alphalyse A/S, Unsbjergvej 4, Odense, Denmark.

Email correspondence: mortz@alphalyse.com

Analytical characterization of purified proteins, in particular recombinant proteins for drug discovery and development, requires confirmation of the full length protein sequence. Here we present a powerful mass spectrometry method (MALDI ISD) that provides partial sequence information of the intact protein with up to 20-50 amino acid residues from both the N-terminal and C-terminal in one single analysis. The analysis can thus confirm expression and purification of the full length protein sequence, and detect unexpected truncations and modifications of the termini. The MALDI ISD analysis can also be applied to N-terminally blocked and PEGylated proteins. Alphalyse provides protein analysis services and here we present results obtained on a variety of proteins.

INTRODUCTION

Structural characterization of purified natural and recombinant proteins traditionally includes N-terminal Edman sequencing to confirm 5-15 amino acid residues from the amino terminus. N-terminal sequencing is a requirment according to the ICH Q6B Guideline for characterization of recombinant proteins for clinical testing, and to demonstrate comparability and consistency between cGMP batches. N-terminal Edman sequencing does not work for N-terminally blocked proteins, and the analysis is time-consuming with a cycle time of 40 minutes per amino acid.

An alternative technique, ISD MALDI MS has been developed and demonstrated in several publications for top-down sequencing of peptides and intact proteins (refs 1-5).

In this article, Alphalyse presents results obtained using ISD MALDI mass spectrometry on a range of purified proteins with Mw from 6-80 kDa. MALDI ISD offers several key benefits compared to Edman sequencing:

- Both N- and C-terminal sequences of 20-50 residues can be obtained.
- N-terminally modified and blocked proteins (acetyla ted, pyroglutamate, PEGylated) can be sequenced.
- Data acquisition is fast.
- Very long sequence reads can be obtained with up to 80 residues from one single ISD MALDI mass spectrum.
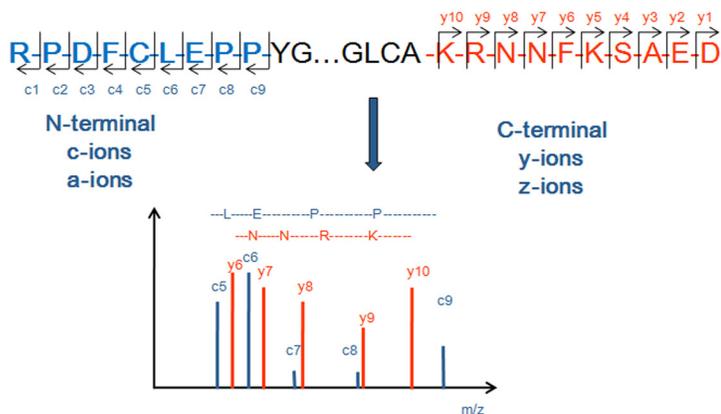
FIGURE 1. PRINCIPLES OF ISD MALDI MS

The intact protein is fragmented in the MALDI ion source. The reflector MALDI spectrum shows intense c-ions (and a-ions) for N-terminal sequencing, as well as y-ions (and z+2 ions) for C-terminal sequencing. Reflector MALDI spectra from 700-5000 Da typically cover 7-50 residues from the terminals.

MATERIALS AND METHODS

Purified proteins were obtained from collaborators and from various commercial suppliers. Two samples were obtained from the ABRF-ESRG 2009 research study. Proteins were reduced with DTT and alkylated with iodoacetamide, and subsequently purified using Biomax-10 Spinfilters or Ziptips (Millipore). The protein samples (20-50 pmol) were deposited on a stainless steel target and co-crystallized with a matrix of 2,5-DHB, 1,5-DAN, or sinapinic acid (refs 1-5). The MALDI mass spectra were acquired on an Autoflex III instrument (Bruker Daltronics) in ISD-mode where the protein ions

spontaneously fragment to yield predominantly c-type fragments with high abundance, and y-, z- and a-ions at lower abundance. The mass spectra were calibrated using external calibration with a standard peptide calibration mixture.

The MS data were correlated to the protein sequences by Mascot database searching (Matrix Science). The Sequence Editor tool in BioTools (Bruker Daltronics) was used for de-novo sequencing and detection of sequence modifications.

RESULTS

A range of different proteins with molecular weights from 6.5 to 77 kDa were analyzed by MALDI ISD. The smallest protein, the aprotinin peptide of 6.5 Da (Figure 2) showed 2 ion-series that confirmed 25 N-terminal

and 23 C-terminal residues. Only 10 amino acids in the middle of the small protein were not covered by the analysis. Thus, 83% of the sequence is confirmed in one single MS analysis.

FIGURE 2 ISD MALDI SPECTRUM OF APROTININ (6.5 kDa)

The MALDI ISD spectrum was used for Mascot database searches (insert) that identified the protein. A c-ion series confirms 25 N-terminal amino acids, and a y-ion series confirms 23 C-terminal residues. The ions and sequences are annotated in the spectrum. residues. Only 10 amino acids in the middle of the small protein were not covered by the analysis. Thus, 83% of the sequence is confirmed in one single MS analysis.
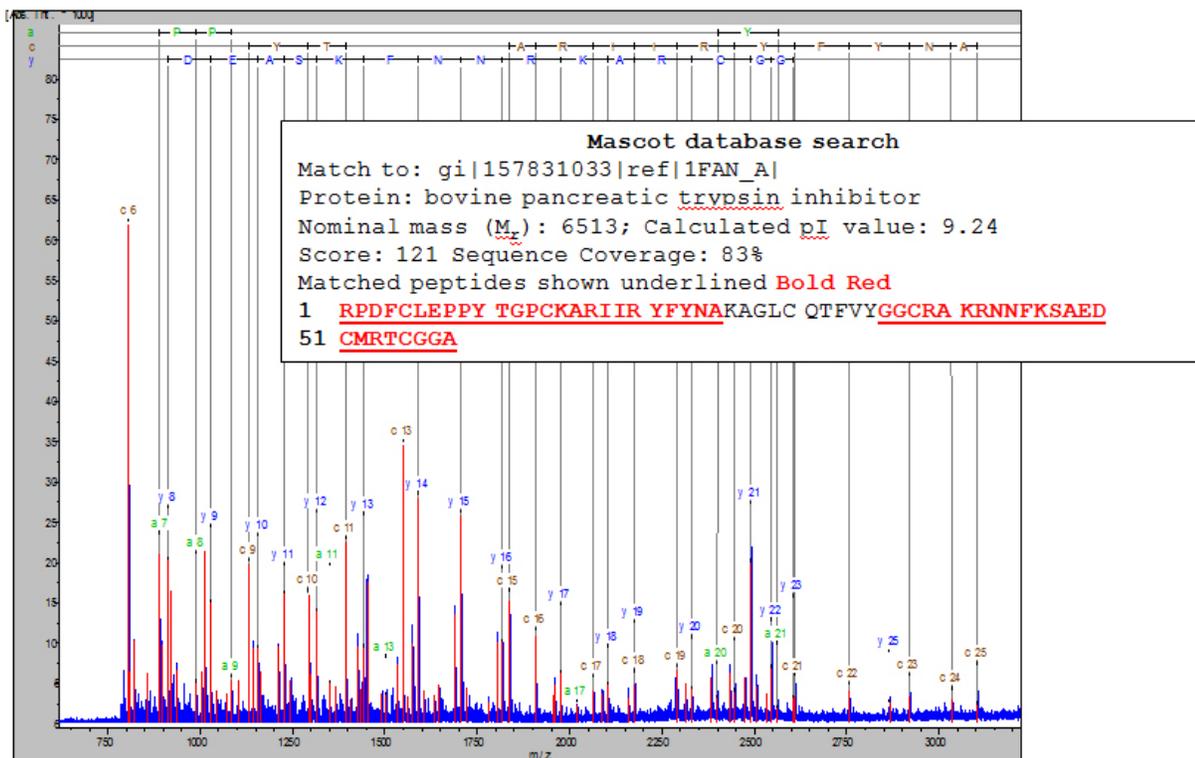
The Association of Biomolecular Resource Facilities (ABRF) 2009 study for comparing Edman and MS based techniques for N-terminal analysis included two protein samples provided to a broad range of laboratories as a blinded study. The 2 protein samples were analyzed by MALDI ISD. The protein in *Sample 1* was identified as a fusion protein between a his-tag expression vector and alcohol dehydrogenase (P00330). The correct protein was identified by a combination of Mascot database searching and *de-novo* sequencing of the fusion area.

The MALDI ISD spectrum covered 30 N-terminal and 29 C-terminal amino acid residues.

The protein in *Sample 2* was identified as GAPDH (P46406). The MALDI ISD spectrum shown in Figure 3 confirmed 39 N-terminal and 32 C-terminal residues. The leading Methionine residue in the database sequence was not present in the N-terminal. See reference 6 for the complete ABRF-ESRG 2009 study results for comparison of Edman and MS techniques for N-terminal sequencing.
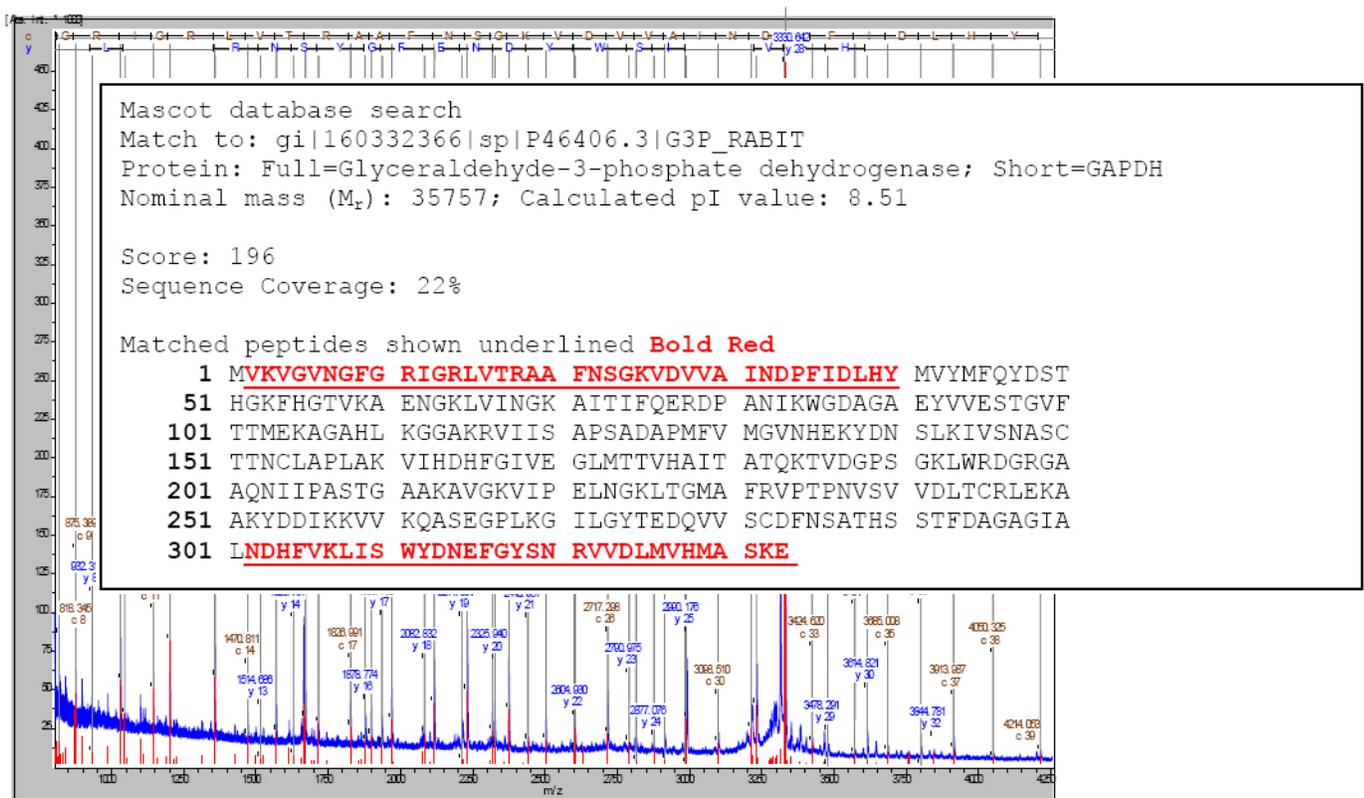


```
Mascot database search
Match to: gi|160332366|sp|P46406.3|G3P_RABIT
Protein: Full=Glyceraldehyde-3-phosphate dehydrogenase; Short=GAPDH
Nominal mass (Mr): 35757; Calculated pI value: 8.51

Score: 196
Sequence Coverage: 22%

Matched peptides shown underlined Bold Red
     1 MVKVGVNGFG RIGRLVTRAA FNSGKVDVVA INDPFIDLHY MVYMFQYDST
    51 HGKFHGTVKA ENGKLVINGK AITIFQERDP ANIKWGDAGA EYVVESTGVF
   101 TTMEKAGAHL KGGAKRVIIS APSADAPMFV MGVNHEKYDN SLKIVSNASC
   151 TTNCLAPLAK VIHDHFGIVE GLMTTVHAIT ATQKTVDGPS GKLWRDGRGA
   201 AQNIIPASTG AAKAVGKVIP ELNGKLTGMA FRVPTPNVSV VDLTCRLEKA
   251 AKYDDIKKVV KQASEGPLKG ILGYTEDQVV SCDFNSATHS STFDAGAGIA
   301 LNDHFVKLIS WYDNEFGYSN RVVDLMVHMA SKE
```

FIGURE 3 ABRF - ESRG STUDY. ANNOTATED ISD MALDI SPECTRUM OF SAMPLE 2 (36 kDa)

The ISD MALDI spectrum was used for a Mascot database search (insert) in the NCBI nrdb database for identification of the protein. The spectrum confirms both the N- and C-terminals of the database sequence, covering in total 71 amino acid residues in a single mass spectrum.

In collaboration with ACE BioSciences A/S, a recombinant vaccine protein was analyzed by MALDI ISD (Figure 4). The Mascot database search confirmed an expected N-terminal truncation where a leader sequence is

removed. The MALDI ISD analysis confirmed the expected C-terminal, and covered 43 N-terminal and 37 C-terminal residues in total.

```
Mascot database search
Match to: gi|86153925|ref|ZP_01072128.1|
Protein: YceI like family protein [Campylobacter jejuni subsp. jejuni HB93-13]
Nominal mass (Mr): 20997; Calculated pI value: 8.83

Score: 263
Sequence Coverage: 47%

Matched peptides shown underlined Bold Red
    1 MKKVLLSSLV AVSLLSTGLF AKEYTLDKAH TDVGFKIKHL QISNVKGNFK
   51 DYSAVIDFDP ASAEFKKLDA TIKIASVNTE NQTRDNHLQQ DDFFKAKKYP
  101 DMTFTMKKYE KIDNEKGKMT GTLTIAGVSK DIVLDTEIGG VAKGKDGKEK
  151 IGFSLNGKIK RSDFKFAAST STITLSDDIN LNIEVEANEK
```
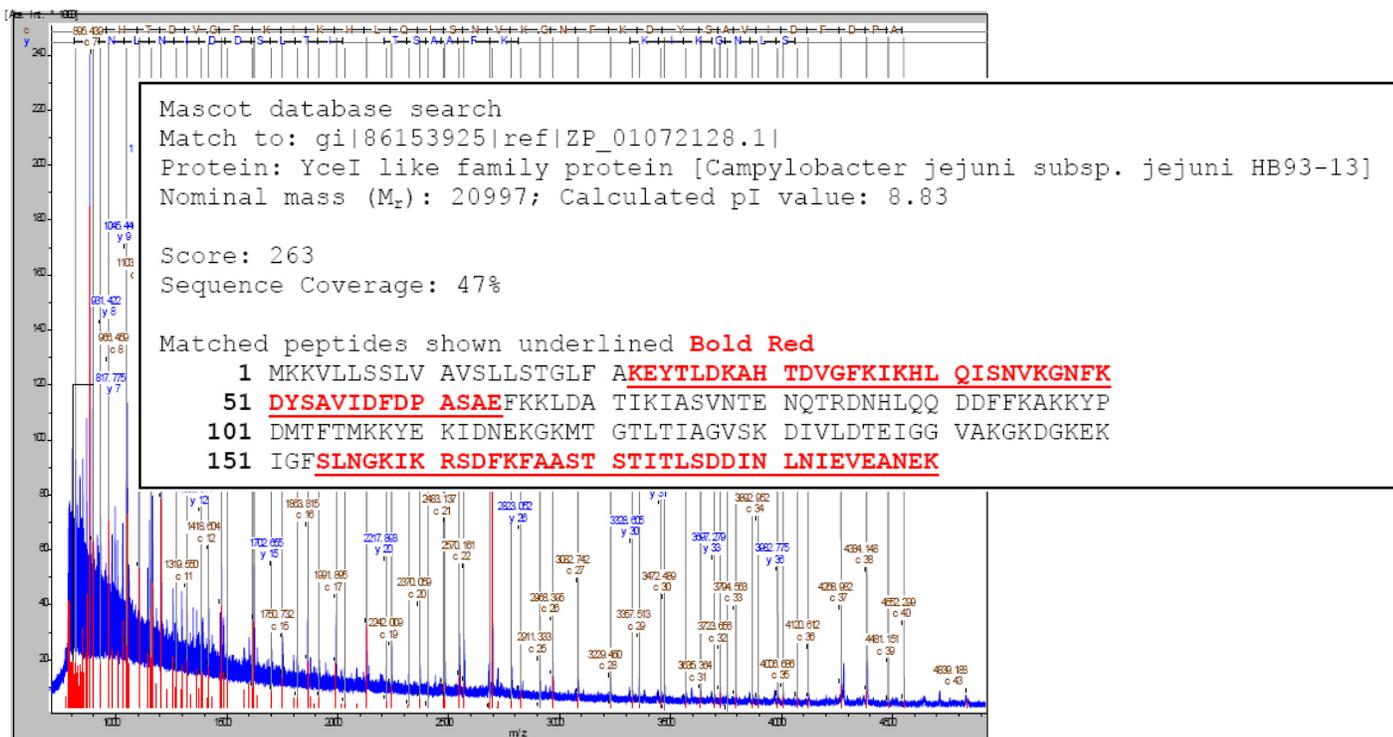
FIGURE 4. ISD MALDI SPECTRUM OF RECOMBINANT VACINE PROTEIN (21 kDa)

The ISD MALDI spectrum and Mascot database search (insert) confirms the C-terminal of the database sequence, and demonstrates an N-terminal truncation. The analysis confirms in total 80 residues in a single ISD MALDI spectrum.

The range of different proteins analyzed by MALDI ISD for this application note is summarized below in Table 1. The analysis was successful for proteins from 6 to 80 kDa. For all proteins analyzed both N-terminal and C-terminal sequences were obtained with sequence lengths from 19-43 residues. In several proteins N-terminal truncations were detected. For the monoclonal antibody, the analysis confirmed the expected N- and C-terminals for both the heavy and the light chains. An N-terminal pyroglutamate residue was detected in the heavy chain.

TABLE 1. LIST OF PROTEINS ANALYZED BY MALDI ISD

PROTEIN NAME	MW [kDa]	N-TERM	C-TERM	COMMENTS
Aprotinin	6.5	25 AA	23 AA	
Myoglobin	16.9	33 AA	31 AA	
ACE393	18.8	43 AA	37 AA	N-terminal truncation
GAPDH	35.8	21 AA	24 AA	
BSA	69.3	33 AA	27 AA	N-terminal truncation
Transferrin	76.9	43 AA	19 AA	
ABRF study 2009: His-tagged ADH	40.1	30 AA	29 AA	
ABRF study 2009: GAPDH	35.8	39 AA	32 AA	N-terminal Met absent
Anti hOX40 light chain	24.2	34 AA	29 AA	
Anti hOX40 heavy chain	48.7	21 AA	22 AA	N-terminal pyroglutamate

CONCLUSION

In the present study, the MALDI ISD technique showed several advantages compared to traditional Edman sequencing. A comparison of the 2 techniques is given in Table 2. The features and advantages of MALDI ISD make it a very useful technique for analysis and quality control of purified natural and recombinant proteins.

- MALDI ISD worked efficiently for all analyzed proteins from 6-80 kDa.
- Both N-terminal and C-terminal sequences were obtained for all proteins.
- Long sequence reads of 19-43 residues were obtained from both termini.
- Truncated and N-terminally blocked proteins could be sequenced.

TABLE 2. COMPARISON OF MALDI ISD AND EDMAN TECHNIQUE FOR PROTEIN SEQUENCING

TECHNIQUE	EDMAN	MALDI ISD
N-terminal sequence	Yes	Yes
C-terminal sequence	No	Yes
Sequence lengths	5-50 residues	20-50 residues
Sample requirements	Purified protein. SDS PAGE purified protein on PVDF membrane	Purified protein
Disadvantage	• Low throughput analysis with cycle time 40 mins per residue. • Expensive chemicals results in high cost per residue. • Does not work for N-terminally blocked proteins. • No information about the C-terminal	• Due to MALDI matrix ions, the first 5-7 amino acids can only be sequenced directly on instruments with "T3-sequencing" utility. • De-novo sequencing of novel proteins is difficult. • Isobaric amino acids cannot be distinguished (I/L, Q/K).
Advantage	• Can be combined with 1D and 2D electrophoresis for analysis of proteins in mixtures. • True sequencing of individual aa makes de-novo sequencing possible.	• C-terminal sequencing possible. • Rapid analysis. • Modified proteins and N-terminally blocked proteins can be sequenced.

REFERENCES

1. Reiber, DC et al. Anal. Chem.. 1998, 70, 673-683. Identifying proteins using matrix- assisted laser desorption/ionization in-source fragmentation data combined with database searching.
2. Takayama, M et al. Electrophoresis 2000, 21, 1670-1677. Sequence information of peptides and proteins with in-source decay in matrix assisted laser desorption/ionization-time of flight mass spectrometry.
3. Schnaible, V et al. Anal. Chem. 2002, 74, 4980-4988. Screening for disulfide bonds in proteins by MALDI in-source decay and LIFT-TOF/TOF-MS.
4. Demeure, K et al. Anal. Chem. 2007, 79, 8678-85. Rational selection of the optimum MALDI matrix for top-down proteomics by in-source decay.
5. Suckau, D et al. Bruker Daltronics Application Notes #MT-57. Reflector in-source-decay (ISD) MALDI-TOF MS: A powerful tool for N-terminal sequence characterization of proteins.
6. Thoma, RS et al. The ABRF ESRG 2009 Study: Comparison of Edman and mass spectrometry techniques for N-terminal sequencing.