ASMS 2024 THP 755

Evaluating a new Workflow for Top-Down Protein Sequence Confirmation and *de novo* Sequencing

Mariangela Kosmopoulou[1]; George Alevizos[1]; Georgia Orfanoudaki[1]; Athanasios Smyrnakis[1]; Michael Greig[2]; Detlev Suckau[3], [1]Fasmatech, Athens, Greece; [2]Bruker Scientific, Billerica, MA; [3]Bruker Daltonics, Bremen, Germany

Introduction

- Top-Down MS sequencing efficiently characterizes proteins N- and C-terminal sequences and assists with PTM localization.
- The complexity of high-resolution MS data limited ESI-Top-Down (TD) analyses to terminal sequence confirmation.
- Thus, the *de novo* determination of protein sequences requires new data analysis software to reliably extract and timely process such high-density information.
- Here, we describe a new software – OmniScape™ - with workflows for protein *de novo* sequencing and sequence confirmation using complex ESI and MALDI TD MS/MS data Fig 1.
- Maximized sequence coverage can be acquired from multiple experiments, while the identification of the best matching proteoforms is another crucial feature.

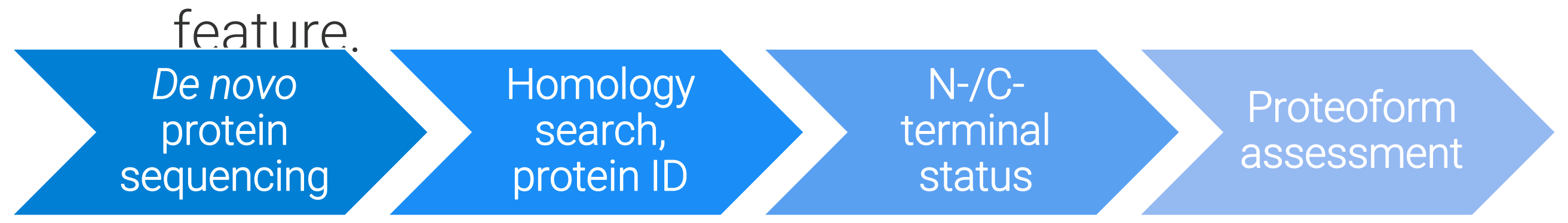

De novo protein sequencing → Homology search, protein ID → N-/C-terminal status → Proteoform assessment

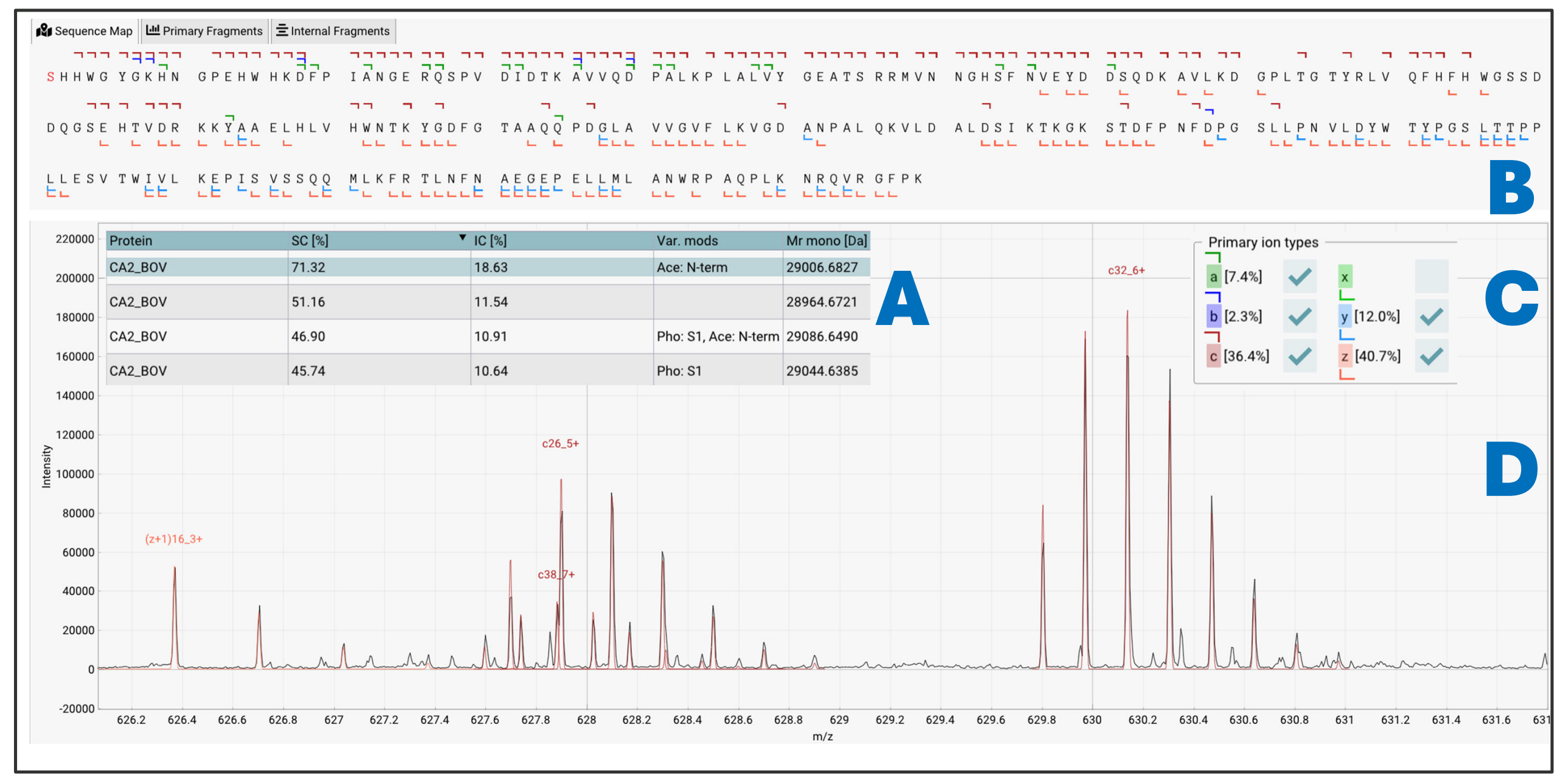
Fig. 1 Scheme of the workflow for seamless proteoform discovery and detailed characterization

Methods

Top-Down ESI-QTOF-MS spectra were acquired after ETD, ECD and EID fragmentation of protein precursor ions.

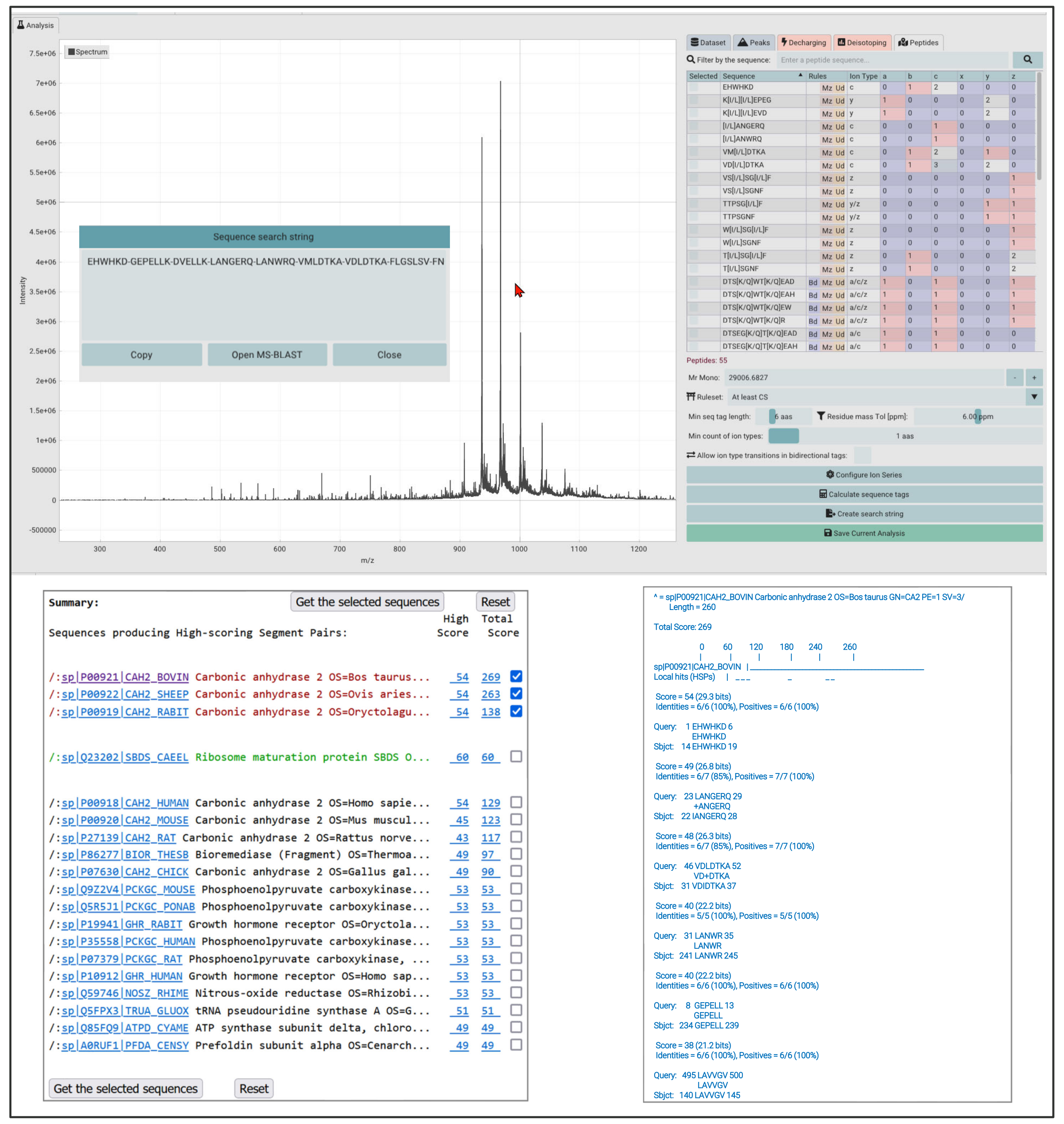
A new analysis software - OmniScape - was developed 1) for the confirmation and spectral annotation of known protein lead sequences, and 2) for the *de novo* sequencing and homology searching of unknown proteins. First, the core algorithm - OmniWave™ - exhaustively extracts possible fragment ion masses, including all charge states and isotopes. In a second step, it determines all possible monoisotopic masses of fragment ions and matches all plausible sequences.

Results


Sequence Map
Primary Fragments
Internal Fragments
SHHWG YGKHN GPEHW HKDFP IANGE RQSPV DIDTK AVVQD PALKP LALVY GEATS RRMVN NGHSF NVEYD DSQDK AVLKD GPLTG TYRLV QFHFH WGSSD
DQGSE HTVDR KKYAA ELHLV HWNTK YGDFG TAAQQ PDGLA VVGVF LKVGD ANPAL QKVLD ALDSI KTKGK STDFP NFDPG SLLPN VLDYW TYPGS LTTPP
LLESV TWIVL KEPIS VSSQQ MLKFR TLNFN AEGEP ELLML ANWRP AQPLK NRQVR GFPK
B
A
C
D
Primary ion types
a [7.4%]
b [2.3%]
c [36.4%]
x
y [12.0%]
z [40.7%]
c32_6+
c26_5+
c38_7+
(z+1)16_3+
Intensity
m/z

Protein	SC [%]	IC [%]	Var. mods	Mr mono [Da]
CA2_BOV	71.32	18.63	Ace: N-term	29006.6827
CA2_BOV	51.16	11.54		28964.6721
CA2_BOV	46.90	10.91	Pho: S1, Ace: N-term	29086.6490
CA2_BOV	45.74	10.64	Pho: S1	29044.6385

Fig. 2 User interface essentials: A) scored list of proteoforms, highest Sequence Coverage (SC) for N-acetyl CAH2, B) Sequence Map with matching fragments which are described in legend C), D) zoomed in view of the ETD spectrum (maXis II ETD) with annotated fragment isotope patterns.



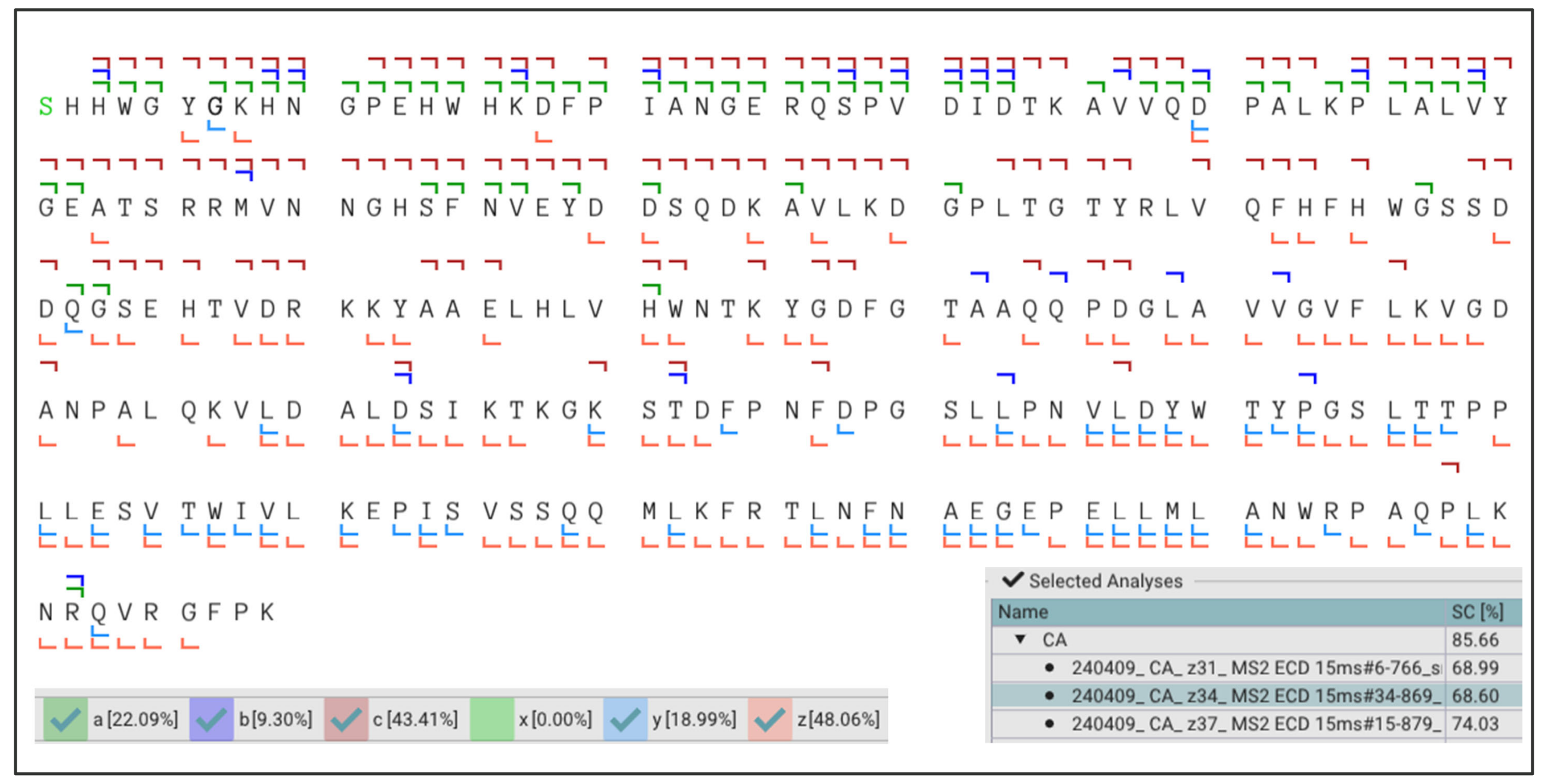
Sequence search string

EHWHKD-GEPELLK-DVELLK-LANGERQ-LANWRQ-VMLDTKA-VDLDTKA-FLGSLSV-FN

```
Summary:                                   Get the selected sequences   Reset
                                                                 High  Total
Sequences producing High-scoring Segment Pairs:                  Score  Score

/:sp|P00921|CAH2_BOVIN Carbonic anhydrase 2 OS=Bos taurus...      54   269  ☑
/:sp|P00922|CAH2_SHEEP Carbonic anhydrase 2 OS=Ovis aries...      54   263  ☑
/:sp|P00919|CAH2_RABIT Carbonic anhydrase 2 OS=Oryctolagu...      54   138  ☑

/:sp|Q23202|SBDS_CAEEL Ribosome maturation protein SBDS O...      60   60   ☐

/:sp|P00918|CAH2_HUMAN Carbonic anhydrase 2 OS=Homo sapie...      54   129  ☐
/:sp|P00920|CAH2_MOUSE Carbonic anhydrase 2 OS=Mus muscul...      45   123  ☐
/:sp|P27139|CAH2_RAT Carbonic anhydrase 2 OS=Rattus norve...      43   117  ☐
/:sp|P86277|BIOR_THESB Bioremediase (Fragment) OS=Thermoa...      49   97   ☐
/:sp|P07630|CAH2_CHICK Carbonic anhydrase 2 OS=Gallus gal...      49   90   ☐
/:sp|Q9Z2V4|PCKGC_MOUSE Phosphoenolpyruvate carboxykinase...      53   53   ☐
/:sp|Q5R5J1|PCKGC_PONAB Phosphoenolpyruvate carboxykinase...      53   53   ☐
/:sp|P19941|GHR_RABIT Growth hormone receptor OS=Oryctola...      53   53   ☐
/:sp|P35558|PCKGC_HUMAN Phosphoenolpyruvate carboxykinase...      53   53   ☐
/:sp|P07379|PCKGC_RAT Phosphoenolpyruvate carboxykinase, ...      53   53   ☐
/:sp|P10912|GHR_HUMAN Growth hormone receptor OS=Homo sap...      53   53   ☐
/:sp|Q59746|NOSZ_RHIME Nitrous-oxide reductase OS=Rhizobi...      53   53   ☐
/:sp|Q5FPX3|TRUA_GLUOX tRNA pseudouridine synthase A OS=G...      51   51   ☐
/:sp|Q85FQ9|ATPD_CYAME ATP synthase subunit delta, chloro...      49   49   ☐
/:sp|A0RUF1|PFDA_CENSY Prefoldin subunit alpha OS=Cenarch...      49   49   ☐

Get the selected sequences   Reset
```

Fig. 3 *De novo* sequence tag generation (top) and MS-BLAST (Harvard, Sunaev lab) homology search result with sequence tags matching to the CAH2 sequence (bottom).

De novo sequence analysis results in several sequence tags that can be used for homology searches in protein sequence databases such as SwissProt. CAH2 was identified from an ETD spectrum in Fig 3.


SHHWG YGKHN GPEHW HKDFP IANGE RQSPV DIDTK AVVQD PALKP LALVY
GEATS RRMVN NGHSF NVEYD DSQDK AVLKD GPLTG TYRLV QFHFH WGSSD
DQGSE HTVDR KKYAA ELHLV HWNTK YGDFG TAAQQ PDGLA VVGVF LKVGD
ANPAL QKVLD ALDSI KTKGK STDFP NFDPG SLLPN VLDYW TYPGS LTTPP
LLESV TWIVL KEPIS VSSQQ MLKFR TLNFN AEGEP ELLML ANWRP AQPLK
NRQVR GFPK
a [22.09%]
b [9.30%]
c [43.41%]
x [0.00%]
y [18.99%]
z [48.06%]

✓ Selected Analyses

Name	SC [%]
▼ CA	85.66
• 240409_CA_z31_MS2 ECD 15ms#6-766_s	68.99
• 240409_CA_z34_MS2 ECD 15ms#34-869_	68.60
• 240409_CA_z37_MS2 ECD 15ms#15-879_	74.03

Fig. 4 Sequence map of CAH2 resulting from the combined maps obtained by ECD analysis of the 31+, 34+ and 37+ precursor ions (timsTOF Omnitrap prototype)

The match between experimental and calculated isotope patterns in the Top-Down ETD spectrum of bovine Carbonic Anhydrase 2 (CAH2) was used for interactive peak list curation. Of the 4 possible proteoforms of CA2 the N-terminally acetylated one scored highest – in agreement with the intact mass – yielding a sequence coverage of 71.3% Fig 2.

In OmniScape, multiple analysis results can be combined into a single sequence map, which increased the sequence coverage of CAH2 from 69% of the individual charge states to 86% for the combined map Fig 4.

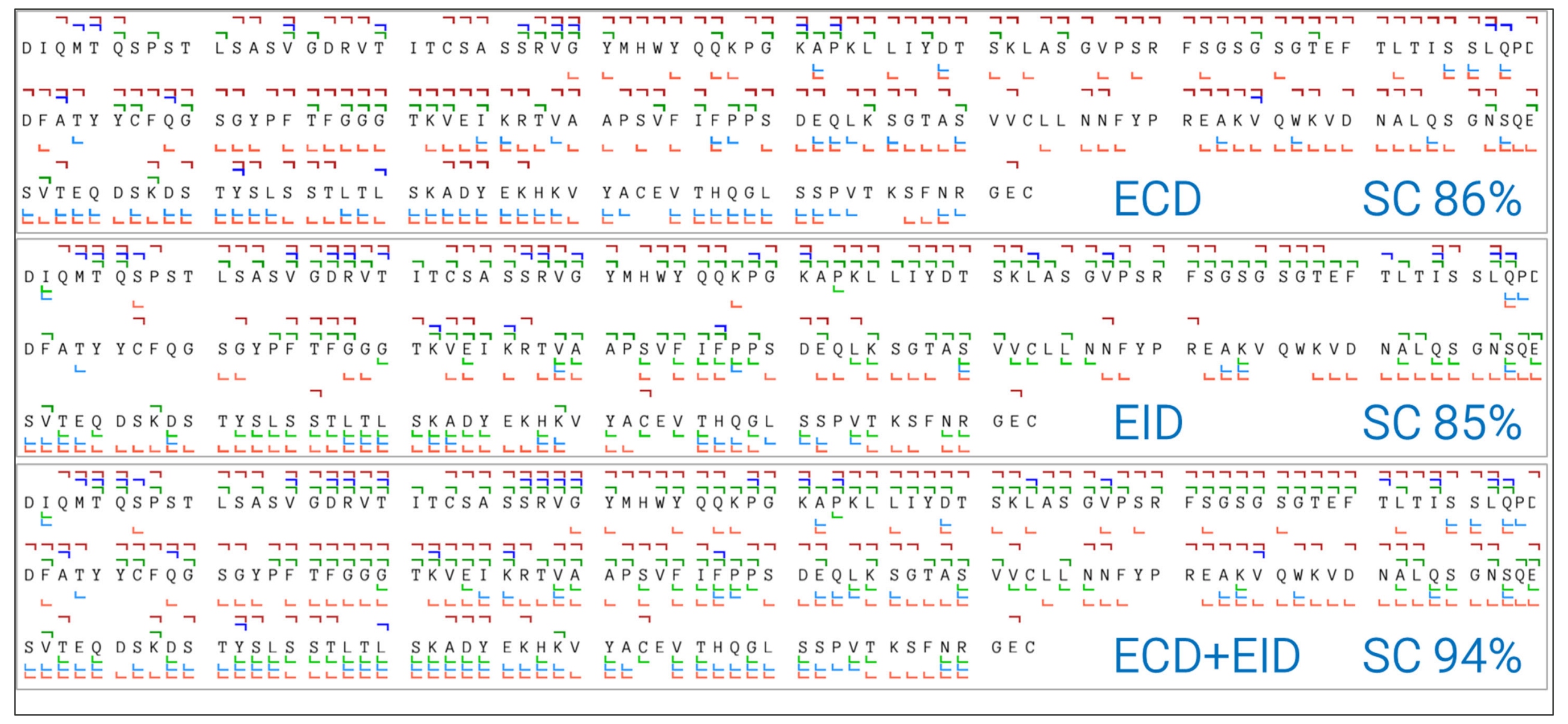

DIQMT QSPST LSASV GDRVT ITCSA SSRVG YMHWY QQKPG KAPKL LIYDT SKLAS GVPSR FSGSG SGTEF TLTIS SLQPE
DFATY YCFQG SGYPF TFGGG TKVEI KRTVA APSVF IFPPS DEQLK SGTAS VVCLL NNFYP REAKV QWKVD NALQS GNSQE
SVTEQ DSKDS TYSLS STLTL SKADY EKHKV YACEV THQGL SSPVT KSFNR GEC
ECD
SC 86%
EID
SC 85%
ECD+EID
SC 94%

Fig. 5 NISTmAb LC subunit analysis with ECD and EID on a timsTOF Omnitrap prototype. Combination of both sequence maps into a combined map increased the sequence coverage (SC) from ~85% each to 94%.

In an analysis of NISTmAb Light Chain different fragmentation methods were combined. Due to the complementarity of EID and ECD fragmentation the sequence coverage could be improved to 94% for the combined map Fig 5.

Summary

The OmniWave algorithm embedded in OmniScape is responsible for the determination of isotope clusters and charge state assignment. In addition, it determines amino acid distances between isotope patterns in the neutral domain to establish sequence tags for the *de novo* sequence determination of a lead sequence Fig 4.

Such a lead sequence can be further analyzed in OmniScape. Proteoforms with various modifications can be scored for the presence of individual structures Fig 2. Even multiple datasets obtained under different conditions or using different fragmentation methods can be combined Fig 5.

Together with the intact neutral mass of a protein of interest, these analyses allow to reliably determine the sequence and its modifications if high resolution/high mass accuracy data are used (here, all analyses used an MS/MS tolerance < 6 ppm) and high-fidelity isotopic clusters intensity is measured by the selected instrument.

The fragment ion representation in OmniScape provides a high information density, enabling the representation of multiple experiment results on limited space Fig 5.

Although high resolution data from ESI-QTOF instruments were discussed here, axial MALDI-ISD data can be reliably processed in OmniScape, as well.

Conclusion

- OmniScape provides full analysis support from Top-Down protein identification through homology searches to proteoform elucidation using a lead sequence
- Multiple datasets across instrument platforms or parameter variation can be combined into a higher sequence coverage representation of the protein
- Powerful interactive data curation in OmniScape safeguards meaningful sequence assessment

Top-Down Protein Sequencing

The authors declare no competing financial interest

For Research Use Only. Not for use in clinical diagnostic procedures.