

High-Throughput de novo Sequencing

Michaela Scigelova¹, Fernando Maroto¹, Craig Dufresne¹, Jesus Vazquez² ● ¹Thermo Finnigan, 355 River Oaks Pkwy. San Jose, CA; ²Centro de Biologia Molecular Severo Ochoa, Madrid, Spain

Overview

De novo sequencing refers to the process of deriving the sequence of an unknown peptide using the information contained in its MS/MS spectrum. The DeNovoX™ software program developed by Thermo Finnigan enables fast and efficient de novo sequencing. The strategy discussed in this presentation uses high confidence sequence tags derived by the program to identify unexpected modifications and amino acid substitutions in complex peptide mixtures.

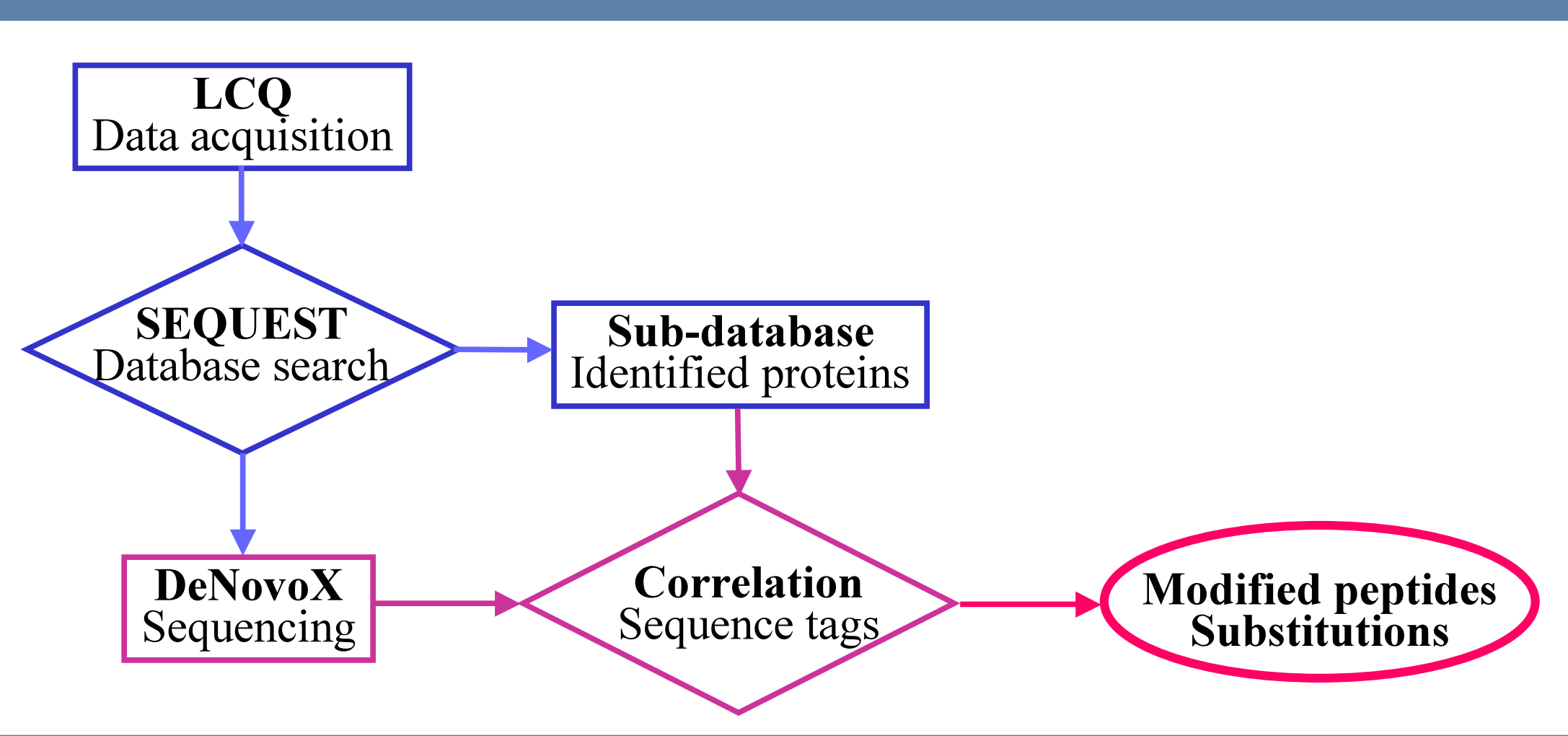
Introduction

Tandem mass spectrometry has become the method of choice for fast and efficient identification of proteins in biological samples. As peptides fragment along the peptide backbone, the MS/MS spectrum contains information about their sequence. The peptide can be identified by a database search only if its sequence had been determined previously and is present in the database. The database search fails if the sequence of the protein is not available or when unexpected modifications and amino acid substitutions are present in real samples.

DeNovoX software is able to derive sequence tags and complete peptide sequences with high confidence from MS/MS spectra acquired in a routine LC-MS/MS experiment. It exceeds the capabilities of a human expert in many instances, and guarantees high-throughput data processing required for proteomics applications.

Although the most obvious use of the program is deriving the peptide sequence for unknown peptides, it can be employed to solve a multitude of other problems. We present an outline of an experiment providing the answer to the detection of unexpected modifications and amino acid substitutions (Figure 1).

FIGURE 1. Outline of the experiment for identification of unexpected modifications and amino acid substitutions



Methods

The protein mixture digested with trypsin was analyzed with the LCQ™ Deca ion trap mass spectrometer (Thermo Finnigan) in a fully automated Data Dependent mode. The data acquired in LC-MS/MS analysis were searched against a public database using SEQUEST®. Proteins identified in this search were used to construct a sub-database used later in the processing.

The spectra not identified by database matching, even though of a good quality, were then batch-sequenced with DeNovoX. The completed and partial sequences (sequence tags) obtained with high confidence by de novo sequencing were then correlated with the peptides in the sub-database constructed previously. This identified peptides with modifications and amino acid substitutions.

Table 1. Nine modified peptides were found in the sample in addition to peptides with carboxyamidomethylated cysteine residues.

(1) K..YICDNQDTISSK..L	BSA	[GY]LZDNQDTLS[SK]	Y → Y+G or Y+57
(2) K..CCTESLVNR...R	BSA	[303]TESLVNR	
(3) K..LGEYGFQNALIVR..Y	BSA	[LGE]YGFQNELLVR	A → E
(4) K..LGEYGFQNALIVR..Y	BSA	[326]EYGFQNALLVR	
(5) SHHWGYGK..H	CA	QHHWGYGK	S → acS (known)
(6) R..QSPVNDTK..A	CA	[198]PVNLDTK	
(7) K..QTALVELLK..H	BSA	[212]ALVELLK	
(8) K..LTFHADICTLPDTEK..Q	BSA	[315]TFHADICTL	L → L+55
(9) R..TLADYNIQK..E	UB	TLSDYNLQK	A → S (known)

Results

LC-MS/MS data from a Thermo Finnigan ion trap contain a wealth of information needed for high confidence de novo sequencing.

Complete sequences and sequence tags obtained with DeNovoX (Figure 2) enable protein identification by correlation to existing databases. In our experiment, 63 peptides were identified by automated batch sequencing (Figure 3).

The strategy outlined in Figure 1 led to the identification of modified peptides and amino acid substitutions (Table 1). Nine modified peptides were found in BSA and carbonic anhydrase. Some of the modifications and substitutions observed (5, 9) had been reported in literature.

A hitherto unknown substitution was confirmed in one of the BSA peptides (3). The peptide (4) appears in three forms – unmodified, and with two different modifications. The acetylated peptide (5) was sequenced for spectra obtained at three different charge states.

In several cases (2, 4, 6, 7) it was possible to pinpoint the location of the modification within two to three amino acid residues. Unambiguous assignment would require MSⁿ experiments on the ion trap.

(SOD) HVGDLGNVTADK
DeNovoX Output for
Tryptic Digest 338.3412.dta

V	83.6%	99.9%
K	83.1%	99.9%
HV	71.1%	99.9%
H	0.6%	99.9%
D	2%	97.7%
DK	68.3%	97.7%
L	74.4%	95.9%
[GD]	69.2%	89.6%
HV[GD]L	46.8%	89.5%
[GD]L	37.2%	89.5%
ADK	48%	82%
A	<0.1%	82%
G	33.6%	71.5%
HV[GD]LG	36.1%	71.5%
T	16.8%	55.5%
TADK	36.2%	53.9%
V	39.7%	40.3%
HV[GD]LGNV	24.4%	39%
VT	21.4%	32%
HV[GD]LGNVT	23.2%	30.9%
VTADT	18.8%	30.9%
LGNVTADT	17.6%	30.9%
V[GD]LGNVTADK	15.9%	29.8%
HV[GD]LGNVTA	17%	29.8%
HV[GD]LGNVTADK	15.8%	28.8%
[GD]LGNVTADK	16.5%	29.8%
HV[GD]LGNVTAD	15.8%	29.8%
HV[GD]LG[GR]TADK	3.6%	13.1%
HV[GD]LGNVADK	2.3%	10%
HV[GD]LG[AN]EADK	0.9%	9.6%

FIGURE 2. The output from DeNovoX stating probabilities for each sequence tag. The completed sequences suggested by the program are in red. Note the consistency among the sequence tags and the first completed sequence (yellow background).

FIGURE 3. Peptides (in orange) identified by de novo sequencing. Unknown amino acid modifications/substitutions highlighted in yellow. Residues highlighted in blue are carboxyamidomethylated cysteines identified by DeNovoX without prior knowledge of this modification.

CA--Carbonic Anhydrase II
SHHWGYGKHGEPZHWHKDFPIANGERQSPVNDTKAVYQDPALKPLALVYGEATSRRMVNN
GHSENVYDSDQDKAVLKDGPLTGTYYRLVQFHFHWGSSBBOQSEHTVDRKKYAAELHLVHWV
TKYGDFTAAQQPDGLAVVGVFLKVGDNALPQKVLDAALDSIKTKGKSTDFNDFGSLLPNV
DYWTYPGSLTTPPLESVTVLKEPISVSSQMLKFRFLNFNAEGEPPELLMLANWRPFAQFLKN
RQVRGFPK

BSA--Bovine Serum Albumin
MKWVTFISLLLFSSAYSRGVFRDRTHKSEIAHRFKDLGEEHFGLVLIASFQYLQCCPFDEHVKL
VNELTEFAKTCVADESHAQCEKSLHTLFGDELKVASLRETYGDMADCEKQEPERNECFLSH
KDDSPDLKLPDPNTLCEFKADEKKFWGKLYEIAARRHPYFYAPELIIYANKYNGVVFQCC
QAEKDKGALLPKIETMREKVLASARQLRCASIQKQGERALKAWSVARLSQKFPKAEFVEVTK
LVTDLTKVHKCECHGDLLECADDRADLAKYCDNQDTISSKKECCDKPLLEKSHCIAEVEK
AIPENLPLTADFAEDKDVCKNYQEAADAFSGFLYEYSRRHPYAVSVLLRLAKEYEATLEEC
AKDDPHACYSTVDFKHLVDEPQNLKQNCQDFEKLGEYGFQNALIVRYTRKVPQVSTPT
LVEVSRSLGKVGTRCCTKPESEMPCTEDYLSLILNRLVLEKTPVSEKVTKCCTESLVNRRP
CFSALTPDETYPKAFDEKLTTFHADICTLPDTEKQIKKQTALVELLKHKPKATEEQLKTVME
NFYAFVDKCCAADDKCAFAVEGPKLVVSTQALA

UB--Ubiquitin
MQIFVKTLTGKTTILEVSSDTIENVKTKIQDKKEGIPDQQLRIFAGKQLEDGRTLADYNIQKES
LHLVLRLRGG

SOD--Superoxide Dismutase
ATKAVCVLKGDPVQGTIHFEAKGDTVVVTSITGLTEGDHGFHVHGFQDNTQGCSTAGPHFN
PLSKKHGGPKDEERHVGDLGNVTADKNGVAIVDIVDPLISLGEYSIIGRTMVVHEKPDLLGRG
GNEESTKTGNAISRACCVIGIAK

Conclusions

The DeNovoX software allows the derivation of sequences for completely unknown peptides.

In addition, it can be used to complement a conventional database search to reveal unexpected modifications, amino acid substitutions or miscleavages.

DeNovoX can perform de novo sequencing with data acquired in a routine LC-MS/MS setup.

A fully automated operation and batch mode sequencing complement the high throughput data analysis delivered by Thermo Finnigan ion trap technology.