Peptide De Novo Sequencing

Peptide *de novo* sequencing

Peptide *de novo* sequencing is the analytical process that derives a peptide's amino acid sequence from its tandem mass spectrum (MS/MS) without the assistance of a sequence database.

It is in contrast to another popular peptide identification approach – "database search", which searches in a given database to find the target peptide.

A clear advantage of *de novo* sequencing is that it works for both database and novel peptides.

Peptide *de novo* sequencing

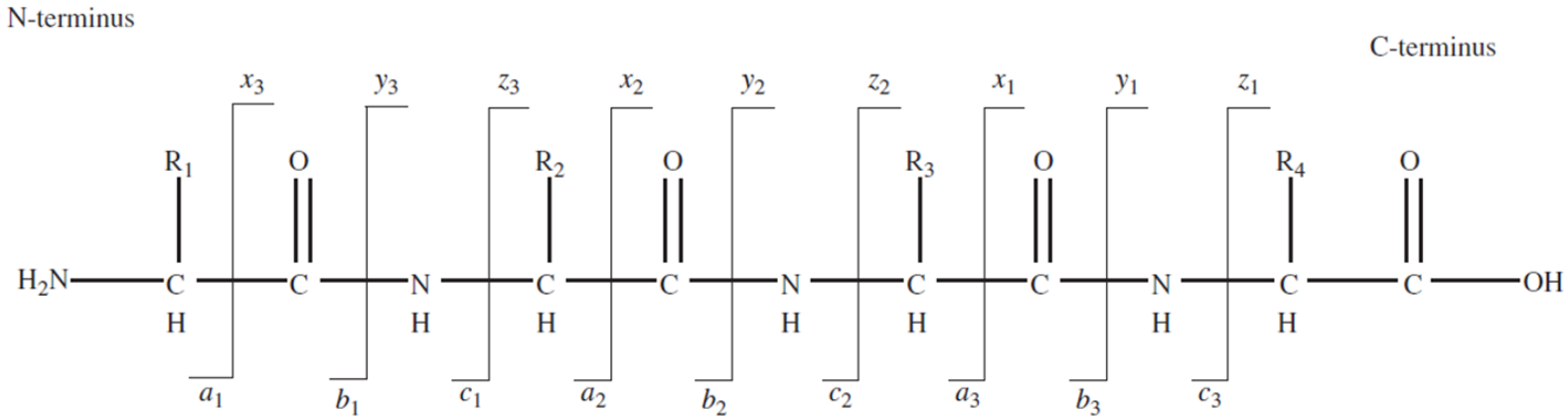
In a tandem mass spectrometer, the peptide is fragmented along the peptide backbone and the resulting fragment ions are measured to produce the MS/MS spectrum. Depending on the fragmentation methods used, different fragment ion types can be produced.

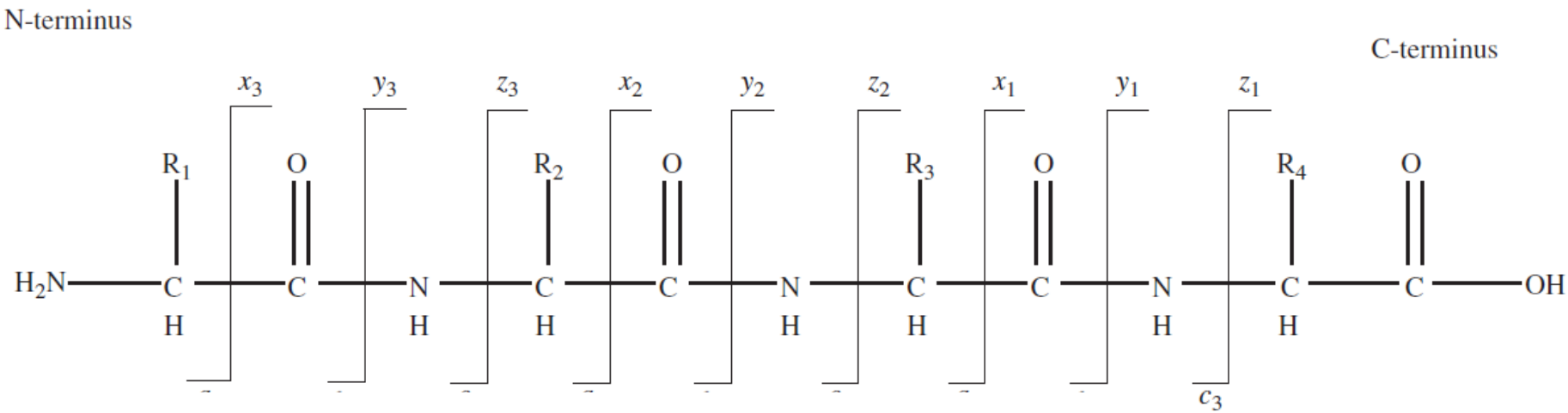
The most widely used fragmentation methods today are Collision-Induced Dissociation (CID) and Electron-Transfer Dissociation (ETD).

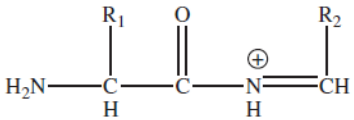
CID produces mostly b and y-ions.

In most mass spectrometers used in proteomic studies the collision energy is considered low (5–50 eV), and the product ions are generally formed through cleavages of the peptide bonds.

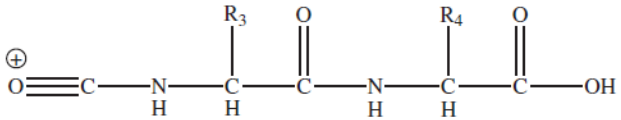
According to the widely accepted nomenclature of Roepstorff and Fohlman, when the charge is retained on the N-terminal portion of the fragmented peptide, the ions are depicted as a, b, and c.
When the charge is retained on the C-terminal portion, the ions are denoted as x, y, and z.



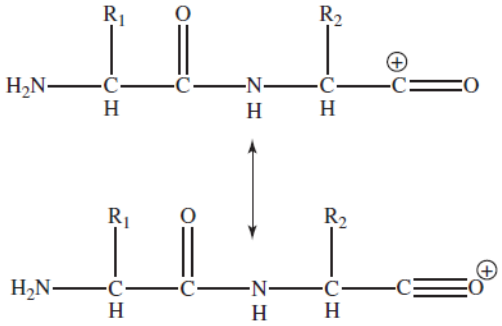




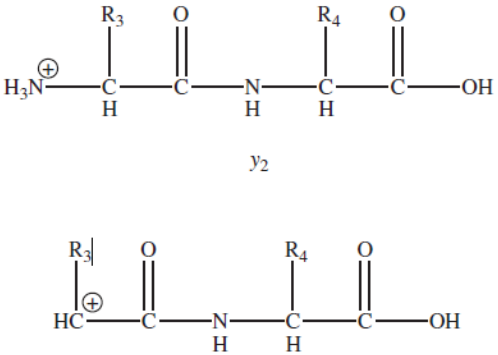
a_2



x_2

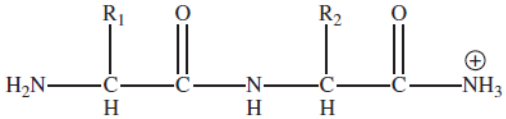


b_2



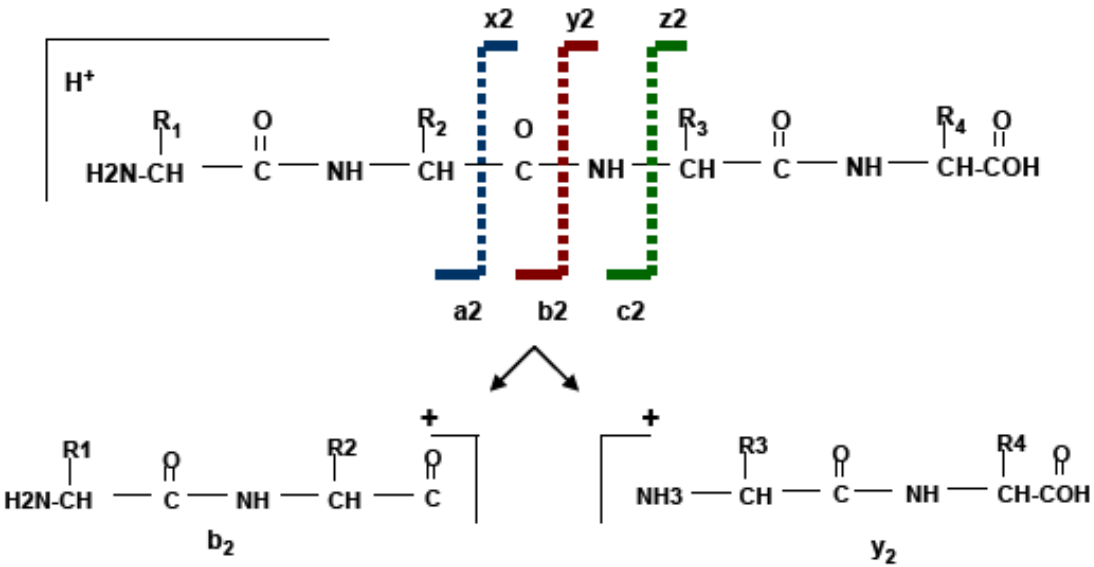
y_2

z_2



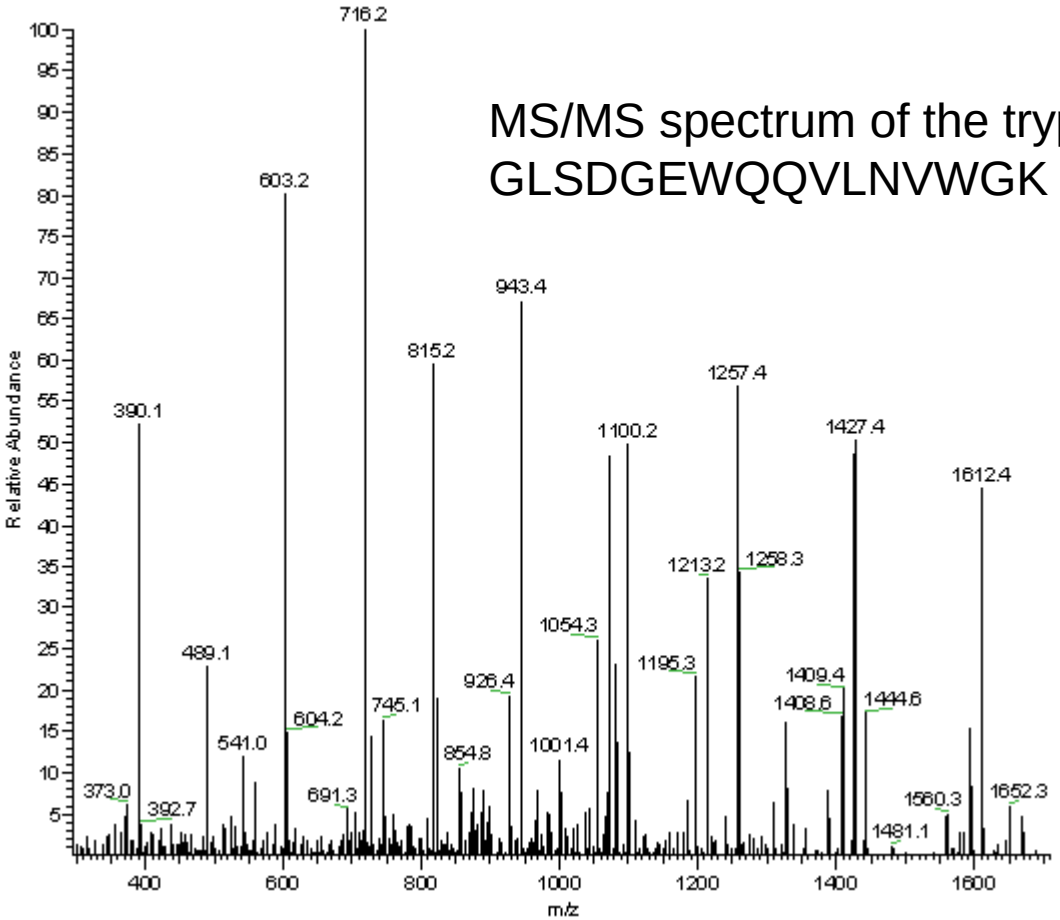
Ion subscript, for example, “2” in y_2, indicates the number of residues (two) contained within the ion

b and y ions



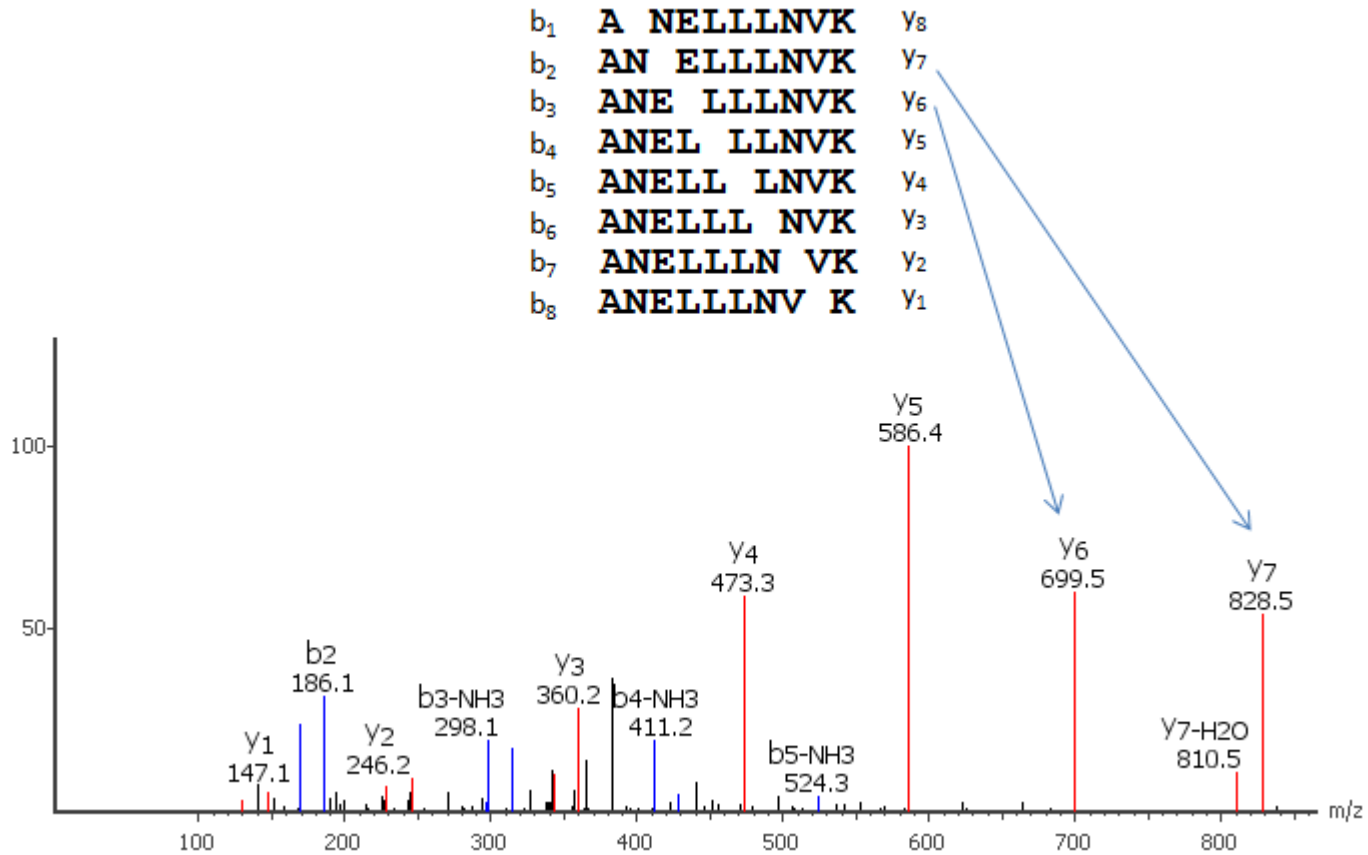
The most common peptide fragments observed in low energy collisions are **a, b** and **y** ions

The **b** ions appear to extend from the amino terminus and **y** ions appear to extend from the carboxyl terminus



In a CID MS/MS, many copies of the same peptide are fragmented at the peptide backbone to form b and y ions. The spectrum consists of peaks at the m/z (mass to charge) values of the corresponding fragment ions. A good quality spectrum often contains many (but not necessarily all) of the theoretical fragment ions.

The main idea of *de novo* sequencing is to use the mass difference between two fragment ions to calculate the mass of an amino acid residue on the peptide backbone. The mass can usually uniquely determine the residue. For example, the mass difference between the y7 and y6 ions is equal to 129, which is the mass of residue E. Similarly, the next adjacent residue between y6 and y5 can be determined as L by the mass difference.



Proteomics 2010, *10*, 634–649

Glycine	Gly	G	57.02146	57.05	C_2H_3NO
Alanine	Ala	A	71.03711	71.08	C_3H_5NO
Serine	Ser	S	87.03203	87.08	$C_3H_5NO_2$
Proline	Pro	P	97.05276	97.12	C_5H_7NO
Valine	Val	V	99.06841	99.13	C_5H_9NO
Threonine	Thr	T	101.04768	101.1	$C_4H_7NO_2$
Cysteine	Cys	C	103.00919	103.1	C_3H_5NOS
Isoleucine	Ile	I	113.08406	113.2	$C_6H_{11}NO$
Leucine	Leu	L	113.08406	113.2	$C_6H_{11}NO$
Asparagine	Asn	N	114.04293	114.1	$C_4H_6N_2O_2$
Aspartic Acid	Asp	D	115.02694	115.1	$C_4H_5NO_3$
Glutamine	Gln	Q	128.05858	128.1	$C_5H_8N_2O_2$
Lysine	Lys	K	128.09496	128.2	$C_6H_{12}N_2O$
Glutamic Acid	Glu	E	129.04259	129.1	$C_5H_7NO_3$
Methionine	Met	M	131.04049	131.2	C_5H_9NOS
Histidine	His	H	137.05891	137.1	$C_6H_7N_3O$
Phenyalanine	Phe	F	147.06841	147.2	C_9H_9NO
Arginine	Arg	R	156.10111	156.2	$C_6H_{12}N_4O$
Tyrosine	Tyr	Y	163.06333	163.2	$C_9H_9NO_2$
Tryptophan	Trp	W	186.07931	186.2	$C_{11}H_{10}N_2O$

Rules

- The **b** ion m/z value is basically the mass of the peptide minus OH, or -17u.
- To calculate the m/z value for the **y** ions just calculate the (M+H)1+ for the shortened peptide

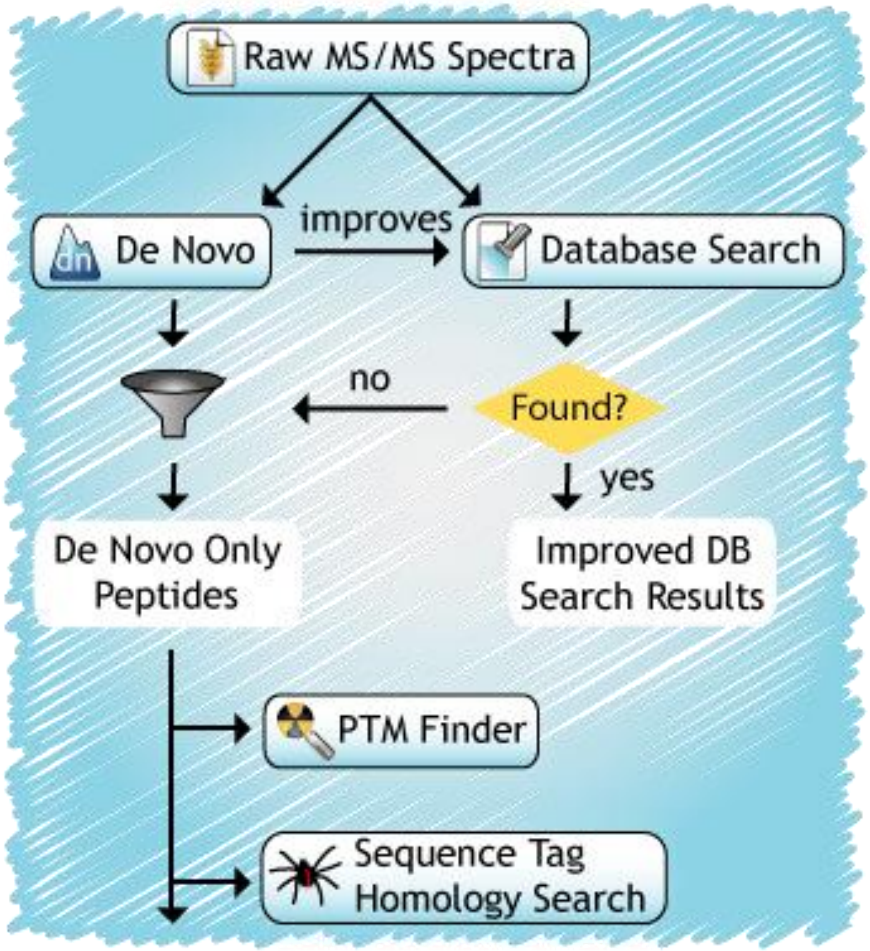
De novo sequencing

Thus, if one can identify either the y-ion or b-ion series in the spectrum, the peptide sequence can be determined.

However, the spectrum obtained from the mass spectrometry instrument does not tell the ion types of the peaks, which require either a human expert or a computer algorithm to figure out during the process of *de novo* sequencing. During this process, a few factors can cause difficulties:

- Incorrect assignment of y and b ions.
- Some fragment ions are missing (such as b1 and y8 in slide 11).
- Existence of other fragment ion types (such as the b3-NH_3 ion).
- Existence of noise peaks in the spectrum.
- The same or similar mass of some residues may cause ambiguity (I=L and K=Q).
- The PTM (post-translational modifications) on the residues may contribute to the mass ambiguity, as well as complicate the peptide fragmentation pattern.

These factors can cause *de novo* sequencing to figure out only a partially correct sequence tag from the spectrum.


Raw MS/MS Spectra
De Novo
improves
Database Search
no
Found?
yes
De Novo Only Peptides
Improved DB Search Results
PTM Finder
Sequence Tag Homology Search

Fragment Ion Calculator

- The calculator takes protein sequences in single-letter code (not including ambiguous amino acids).
- Each sequence should be written on its own line.
- Whitespace and numbers are ignored within the sequence.

Peptide Sequence

Peptide:

GLSDGEWQQVLNVWGK

Mass type: MONO / AVG

Charge state: +1 / +2 / +3

Ion types: A, B, C, X, Y, Z

Submit Reset

Modifications (optional)

Add to N- or C-terminus: N-terminus 0.0 C-terminus 0.0

Add to all AA residues and/or specific location: AA or Pos / Value

e.g. C 57.0
3 80.0
(add +57 to all Cys and
add +80 to 3rd AA residue)

Free on-line calculator provided by the Institute of Systems Biology.

http://db.systemsbiology.net/proteomicsToolkit/FragIonServlet.html

Fragment Ion Calculator Results

Sequence: **GLSDGEWQQVLNVWGK**, pI: **4.37029**

Fragment Ion Table, monoisotopic masses

Seq	#	B	Y	# (+1)
G	1	58.02933	1815.90301	16
L	2	171.11340	1758.88155	15
S	3	258.14543	1645.79749	14
D	4	373.17237	1558.76546	13
G	5	430.19383	1443.73851	12
E	6	559.23642	1386.71705	11
W	7	745.31574	1257.67446	10
Q	8	873.37431	1071.59515	9
Q	9	1001.43289	943.53657	8
V	10	1100.50131	815.47799	7
L	11	1213.58537	716.40958	6
N	12	1327.62830	603.32551	5
V	13	1426.69671	489.28259	4
W	14	1612.77602	390.21417	3
G	15	1669.79749	204.13486	2
K	16	1797.89245	147.11340	1

Mass/Charge Table

	Mass	
	Mono	Avg
(M)	1814.89519	1816.00312
$(M+H)^{+}$	1815.90301	1817.01106
$(M+2H)^{2+}$	908.45544	909.00952

Peptide *De Novo* Sequencing with Peaks

- PEAKS is a tool for *de novo* sequencing in mass spectrometry labs.
- PEAKS assigns a local confidence score for each amino acid in the de novo sequence. This local confidence ranges from 0% to 99%, indicating how confident the algorithm is about the particular amino acid. The whole peptide is evaluated by two measures: the ALC (Average of Local Confidence) and TLC (Total of Local Confidence) scores.

ALC reflects the average correct ratio for the amino acids in the sequence (or the likelihood of each amino acid assignment in a resultant peptide)

TLC reflects the expected total number of correct amino acids in the sequence.

https://www.youtube.com/watch?v=lyhpRu6s7Ro

Summary | Protein | Peptide | De novo only | LC/MS

Show all proteins in each group — **Identified Protein (Groups)** — accession contains search — no results

Accession	-10lgP	Coverage	#Peptides	#Unique	PTM	Avg. Mass	Description	Mark
P06733|ENOA_HUMAN	261.94	74%	28	27	c d o	47169	Alpha-enolase OS=Homo sapiens GN=ENO1 PE=1 SV=2	☑
P13929|ENOB_HUMAN	147.72	22%	7	7	c	46987	Beta-enolase OS=Homo sapiens GN=ENO3 PE=1 SV=5	☑
P09104|ENOG_HUMAN	120.88	17%	5	5	c	47269	Gamma-enolase OS=Homo sapiens GN=ENO2 PE=1 SV=3	☑
Q8NHV1|GIMA7_HUM...	23.79	2%	1	1		34509	GTPase IMAP family member 7 OS=Homo sapiens GN=GIMAP7 PE...	☑
P51149|RAB7A_HUMAN	171.45	74%	14	13	c	23490	Ras-related protein Rab-7a OS=Homo sapiens GN=RAB7A PE=1 S...	☑
P68366|TBA4A_HUMAN	250.84	74%	26	3	c o	49924	Tubulin alpha-4A chain OS=Homo sapiens GN=TUBA4A PE=1 SV=1	☑
P08238|HS90B_HUMAN	307.51	73%	47	26	c d o	83264	Heat shock protein HSP 90-beta OS=Homo sapiens GN=HSP90AB1...	☑
Q58FF7|H90B3_HUMAN	234.98	40%	23	12	c o	68325	Putative heat shock protein HSP 90-beta-3 OS=Homo sapiens GN=...	☑
Q58FF8|H90B2_HUMAN	198.46	33%	14	4	c d o	44349	Putative heat shock protein HSP 90-beta 2 OS=Homo sapiens GN=...	☑

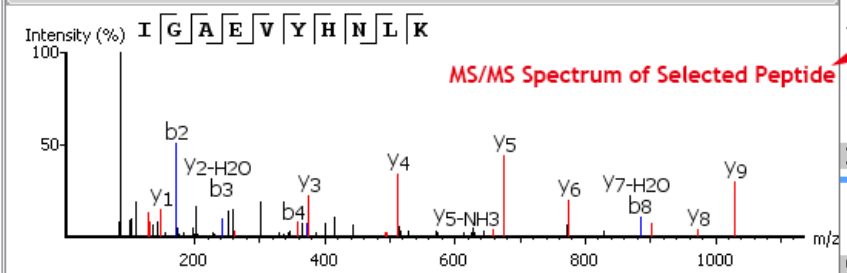
Coverage | Peptides | Denovo Only Tags — **Selected Protein Coverage View**

>sp|P06733|ENOA_HUMAN Alpha-enolase OS=Homo sapiens GN=ENO1 PE=1 SV=2

```
1   MSILKIHARE IFDSRGNPTV EVDLFTSKGL FRAAVPSGAS TGIYEALELR DNDKTRYMGK GVSKAVEHIN

71  KTIAPALVSK KLNVTEQEKI DKLMIEMDGT ENKSKFGANA ILGVSLAVCK AGAVEKGVPL YRHIADLAGN
                                          119 c
                                              c
                                              o

AMRIGAEVYH NLKNVIKEKY GKDATNVGDE
FFRSGKYDLD FKSPDDPSRY ISPDQLADLY
                          339 c
TVTNPKRIAK AVNEKSCNCL LLKVNQIGSV
                          c c

351 TESLQACKLA QANGWGVMVS HRSGETEDTF IADLVVGLCT GQIKTGAPCR SERLAKYNQL LRIEEELGSK
        357 c   363 d
            c                       c
            d

421 AKFAGRNFRN PLAK
```

18 spectra

Sequence=IGAEVYHNLK, Scan F1:10894, m/z=572.3116, z=2, RT=50.05, -10lgP=58.26, ppm=0.2, by ...



MS/MS Spectrum of Selected Peptide

18 spectra; -10lgP=58.26
Enzyme=Trypsin
Seq=IGAEVYHNLK

| 1:1 | 2X | 2Y | ErrTol: 0.05 Da | ☑ preprocess | ☑ low intens. label

outline | coverage

70 AAs per line | ☑ 10AA gap

Sequence Display Option: Fasta

Show de novo only sequences

☐ de novo tags sharing 6 AAs

☐ de novo peptides fully matched

☑ show confident modification site

minimal ion intensity 5 %

minimal Ascore 20.0

	ΔM	PTM	#
☑ c	+57.02	Carbamidomethyla...	29
☑ d	+0.98	Deamidation (NQ)	19
☑ o	+15.99	Oxidation (M)	16

Protein expression comparison among samples

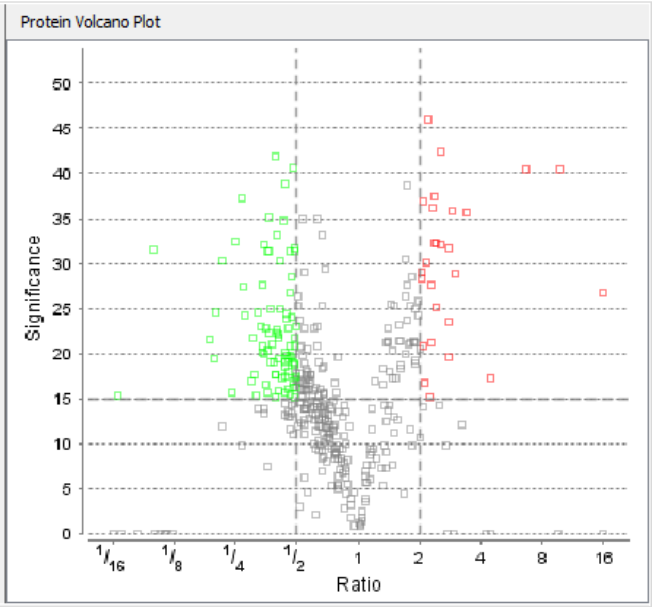
Protein expression comparison between groups

Summary | Protein | Peptide | LC/MS

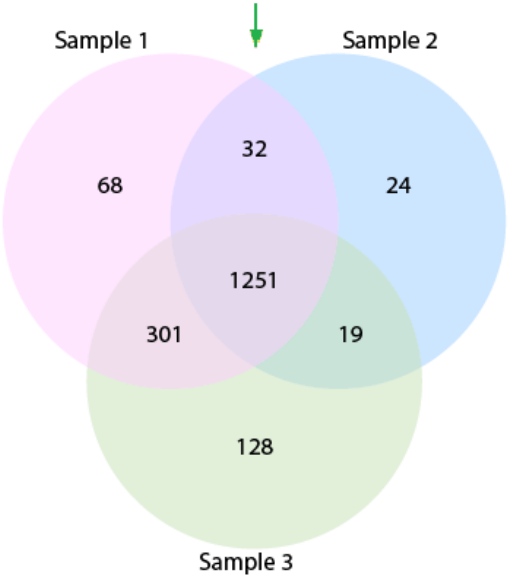
Show top proteins in each group — accession contains search — no results

Accession	Significance	-10lgP	Sample Profile	Group Profile	Group 1_Area	Group 2_Area
P15090|FABP4_HUMAN	57.44	151.22			7.34E6	2.96E8
Q92597|NDRG1_HUMAN	55.99	217.22				4.16E8
P02751|FINC_HUMAN	54.69	340.54				4.59E7
P00352|AL1A1_HUMAN	51.47	240.92				9.93E7
P12277|KCRB_HUMAN	50.64	200.45				4.66E6
P06703|S10A6_HUMAN	49.96	74.89				4.48E5
P08123|CO1A2_HUMAN	48.37	130.19				0
P16401|H15_HUMAN	47.26	170.59				2.34E7
P52943|CRIP2_HUMAN	46.88	160.58				0
P05204|HMGN2_HUMAN	46.39	113.47			7.54E6	0
P06454|PTMA_HUMAN	45.89	153.20			6.9E6	4.88E4
Q9UGM6|SYWM_HUMAN	45.56	65.49			6.03E6	0
Q96B97|SH3K1_HUMAN	45.53	115.33			3.88E6	0
P30837|AL1B1_HUMAN	45.50	112.62			6.92E6	0
P31947|1433S_HUMAN	45.50	192.14			1.68E7	1.31E8
P05787|K2C8_HUMAN	45.31	289.00			8.06E8	8.48E7
Q08380|LG3BP_HUMAN	45.21	140.14			1.75E5	6.72E6
Q9UH65|SWP70_HUMAN	45.10	59.78			2.46E6	0
Q96FW1|OTUB1_HUMAN	45.07	140.62			1.24E7	1.32E5

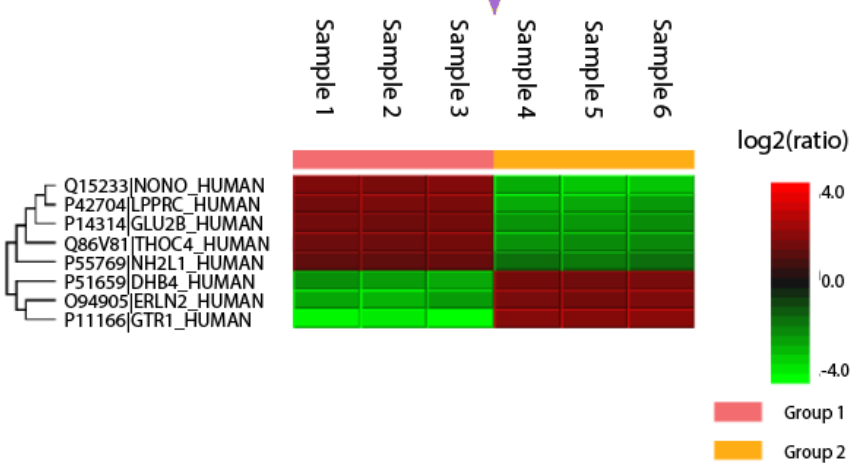
Channel	Area	Ratio
092-1:Light	5.013E6	1.00
092-2:Light	5.836E6	1.16
092-3:Light	1.116E7	2.23
561-1:Light	3.379E8	67.41
561-2:Light	2.603E8	51.92
561-3:Light	2.885E8	57.56



Venn diagram showing overlaps of identified proteins among samples



Protein Profile Heatmap



De Novo

Tools

. Data Refinement

. Replicate Analysis

✓ De Novo

. PEAKS Search

. SPIDER Search

. PTM Finder

De Novo Predefined parameters Instrument_default Save as...

Error Tolerance

Parent ion: 0.1 Da Fragment ion: 0.1 Da

Enzyme

Trypsin View Enzyme New Enzyme...

PTM

Set PTM... Remove Switch type

Maximum allowed variable PTM per peptide 3

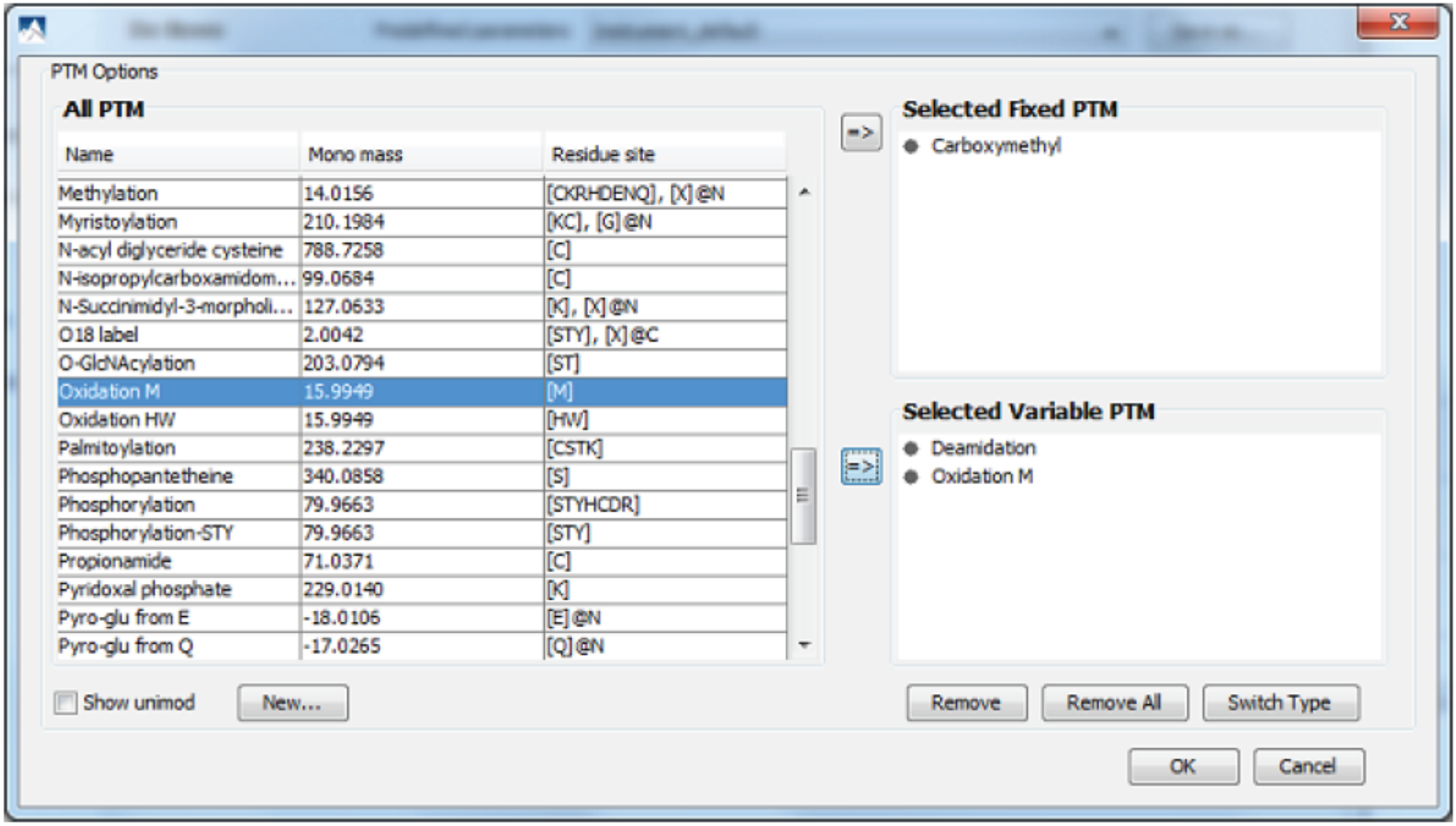
General Options

Report up to 5 candidates per spectrum

OK Cancel Help

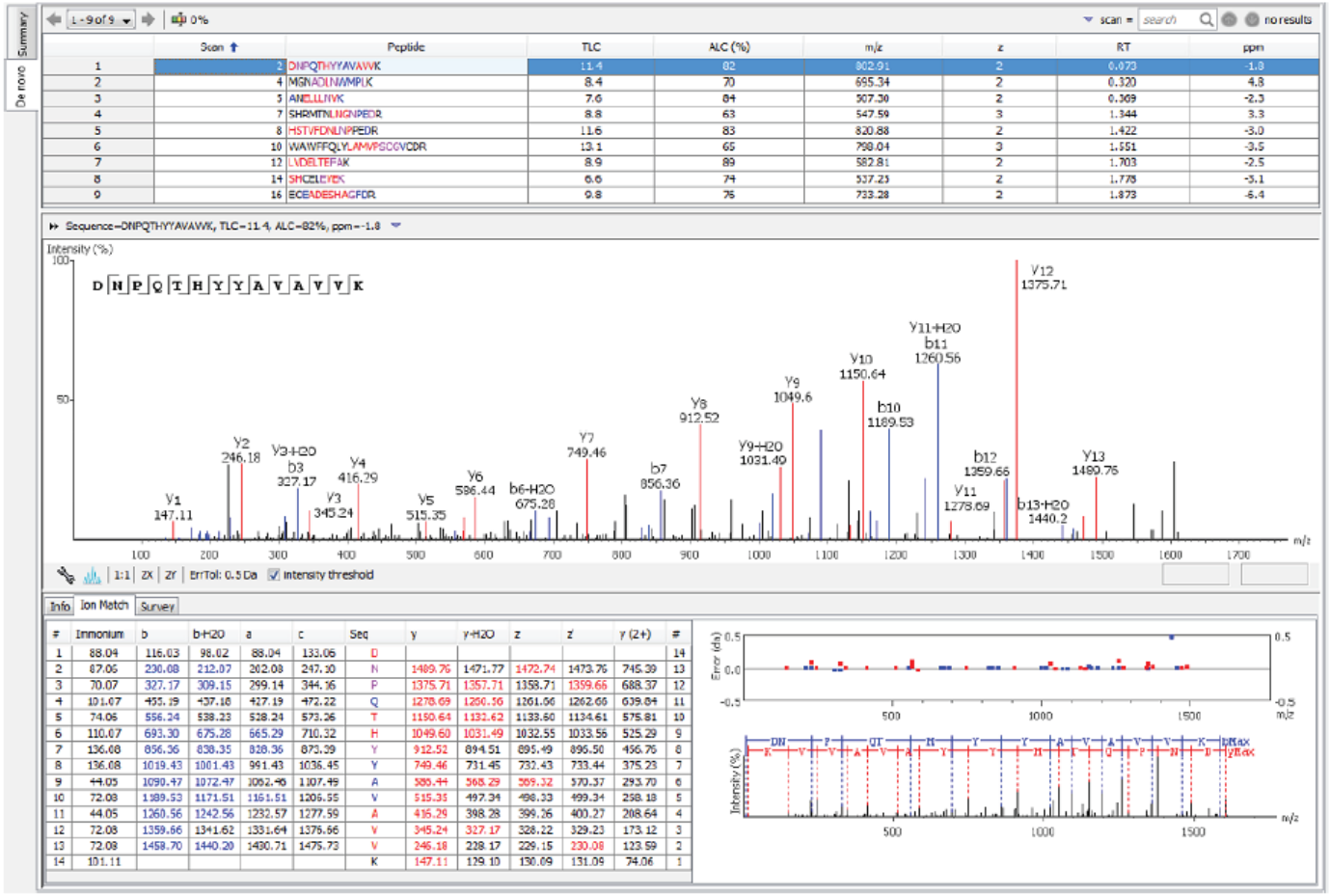
Fixed and Variable PTMs

To select the PTMs for the *de novo* sequencing, click the "Set PTM..." button to open the "PTM Setup" window.



De Novo Peptide View

The *de novo* peptide view displays the *de novo* sequencing results in more detail. The table at the top displays all the *de novo* sequences, and the bottom half of the view provides additional information about the peptide-spectrum match.



Peptide Table

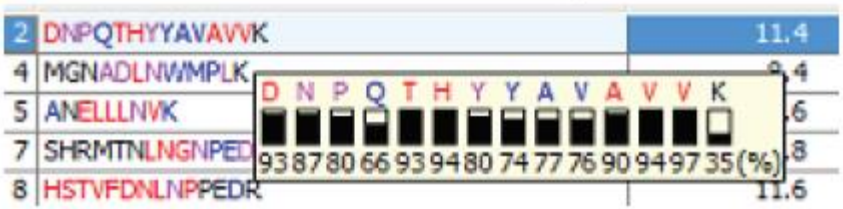
- PEAKS displays the peptide **sequence candidates** at the top of the screen. You can sort the results by clicking on the titles of the columns.

Contents of the columns in the “Peptide Candidates Frame”. The first column is a unique index for the peptides in the list.

- **Scan:** Scan number.
- **Peptide:** The amino acid sequence of the peptide as determined by *de novo* sequencing. If there is any PTM on an amino acid, the amino acid is followed by a pair of parentheses enclosing the delta mass of the PTM.
- **TLC:** Total local confidence. It is calculated by adding the local confidence for each amino acid in the peptide sequence.
- **ALC:** Average local confidence (TLC divided by the peptide length).
- **m/z:** The measured mass/charge value, in Daltons, for the spectrum.
- **z:** The calculated charge value for the peptide.
- **RT:** Retention time (elution time) for the spectrum as recorded in the data.
- **ppm:** The precursor mass error, calculated as 106 × (observed mass - theoretical mass) / theoretical mass.

Confidence Scores

- Next to the proposed sequence candidates, the auto *de novo* “Total Local Confidence” (TLC) and “Average Local Confidence” (ALC) confidence scores are shown.
- The local confidence scores for each amino acid (that is, confidence that the correct residue in each position has been identified) are represented by color coding.
- Red represents a very high confidence (greater than 90%), purple represents a high confidence (80 to 90%) blue represents a medium confidence (60 to 80%) and black represents a low confidence (less than 60%).



Spectrum Annotation