Proteomics

BIO390 “Introduction to Bioinformatics”

29.10.2019

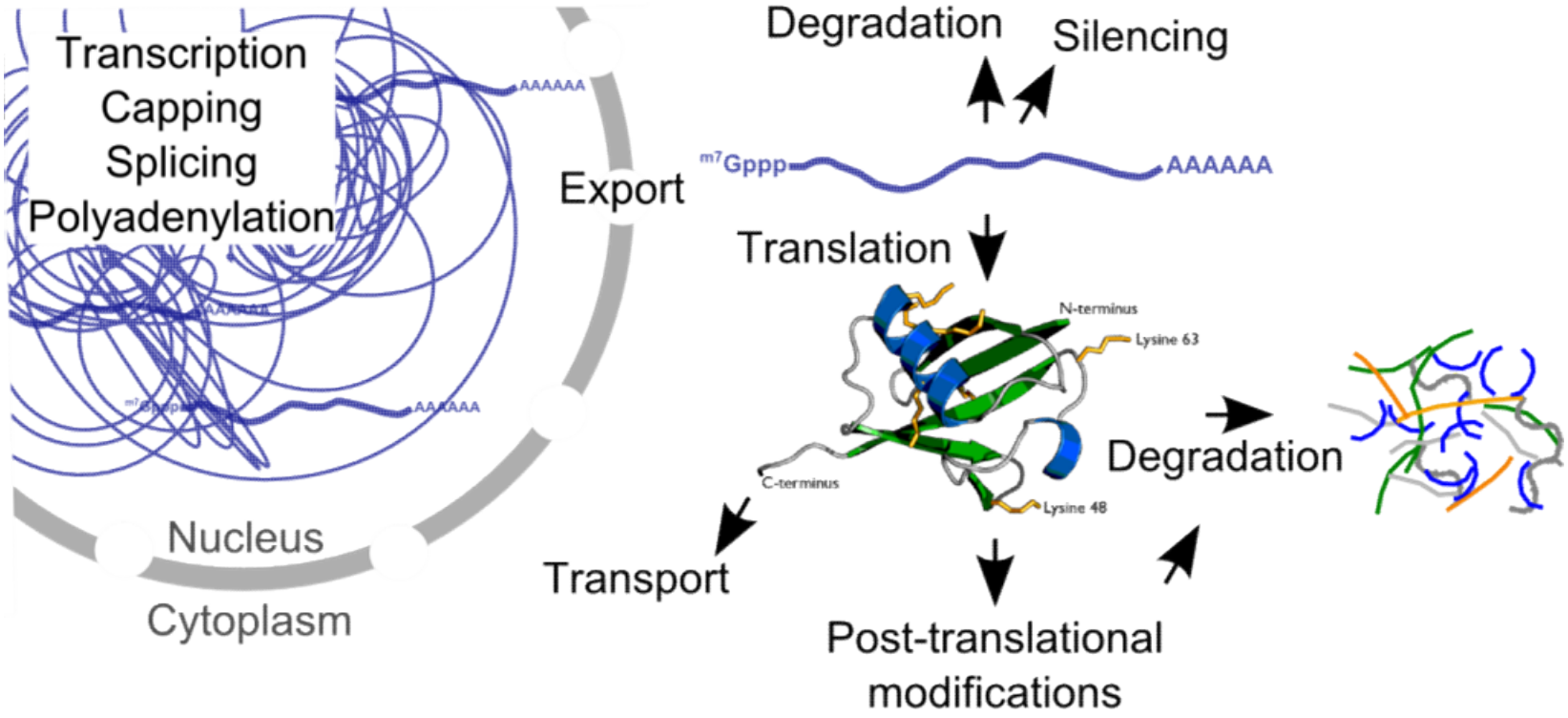
PD Dr. Katja Bärenfaller
katja.baerenfaller@siaf.uzh.ch

In proteomics one of the important bioinformatics tasks is to generate lists of reliably identified peptides and proteins in mass spectrometry-based experiments. For this, amino acid sequences are assigned to measured tandem mass spectra. The quality of the peptide spectrum assignments are scored and criteria are applied that allow to distinguish the good from the bad hits and to estimate the quality of the dataset.

In the context of this lecture, you will need to learn and understand:

- what information can be gained in a proteomics experiment
- what's the principle of assigning an amino acid sequence to a tandem mass spectrum
- how *de novo* and database-dependent peptide identifications work
- one way of how the accuracy of peptide identifications can be estimated
- how the number of wrong hits in a dataset can be estimated
- current proteomic approaches

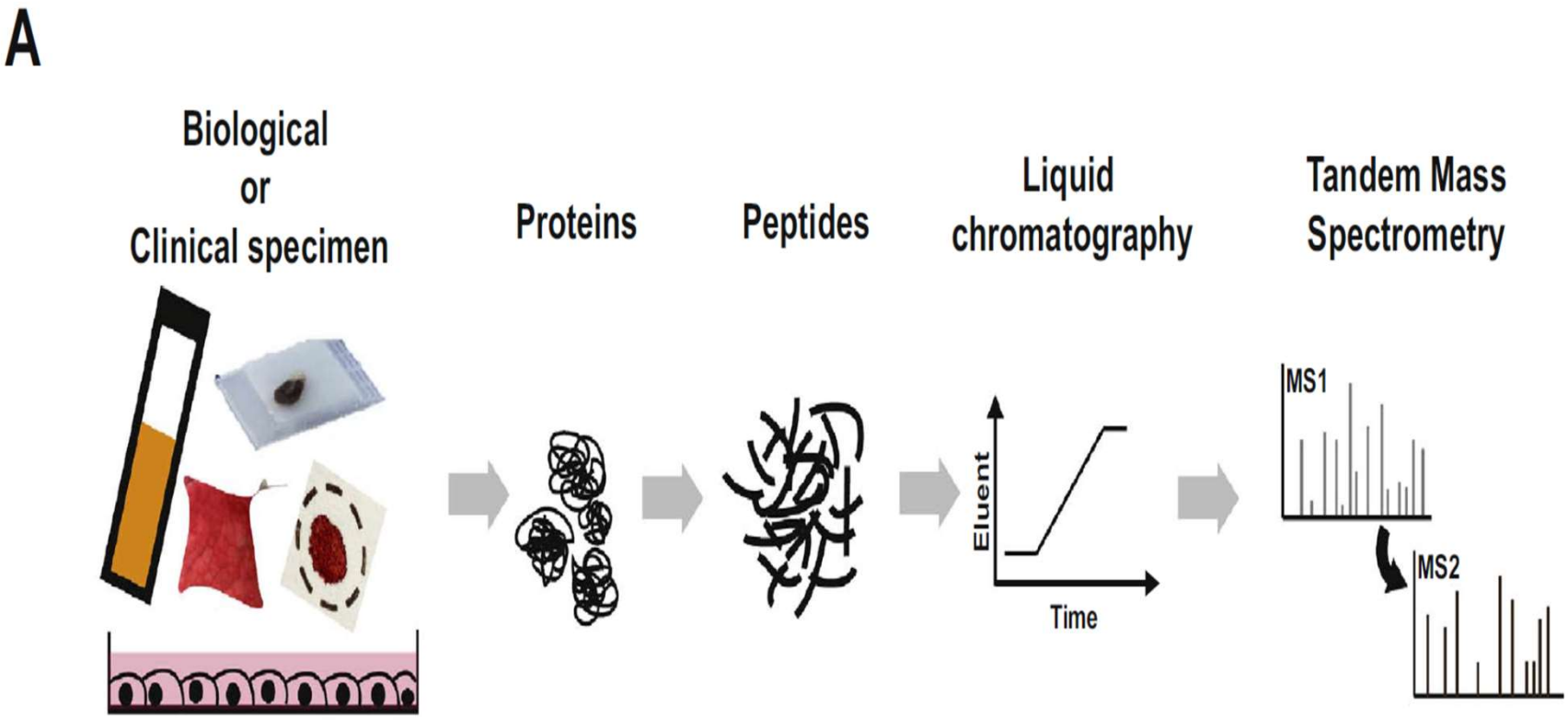
Various processes determine protein levels and activities



→ Not only the genome, but in particular the proteins present and their activities, their sub-cellular localisation, and their protein-protein or protein-DNA/RNA interactions determine the appeareance and state of a biological organism

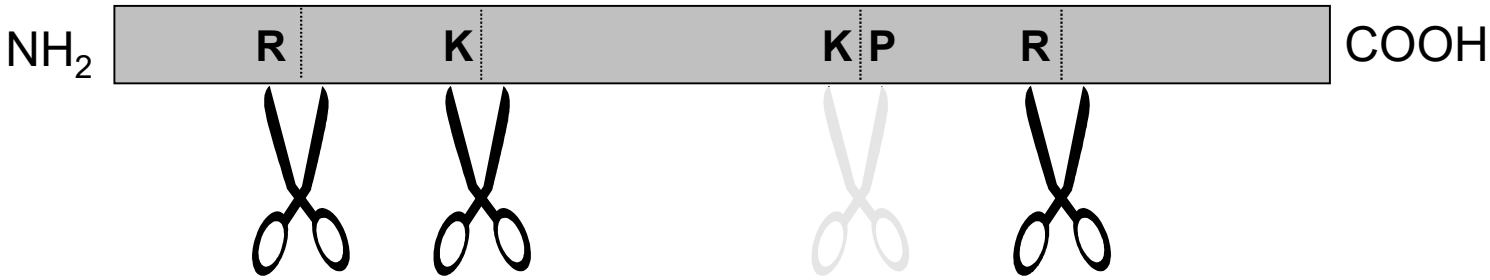
→ Gene expression is regulated on many different levels, including also enhanced or reduced translational efficiency, increased or decreased protein degradation, triggering of signaling cascades, e.g. through protein phosphorylation

Generic mass spectrometry-based proteomics experiment

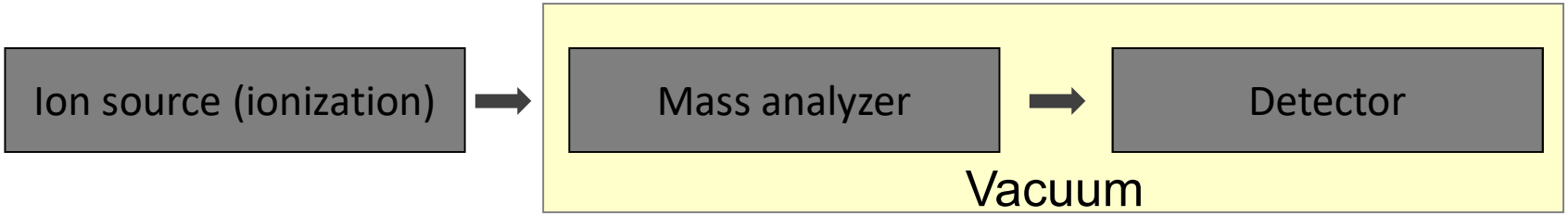


Uzozie & Aebersold, Journal of Proteomics, 2018

- Before analysis, the proteins are typically digested with a site specific protease, most of the time with trypsin. Trypsin cuts after arginine or lysine, except when the cutting site is followed by proline, which leads to limited cleavage.



The components of a mass spectrometer



ESI = **E**lectro**s**pray **I**onisation

MALDI = **M**atrix **A**ssisted **L**aser **D**esorption **I**onisation

TOF = **T**ime **o**f **F**light

Quadrupole (Q)

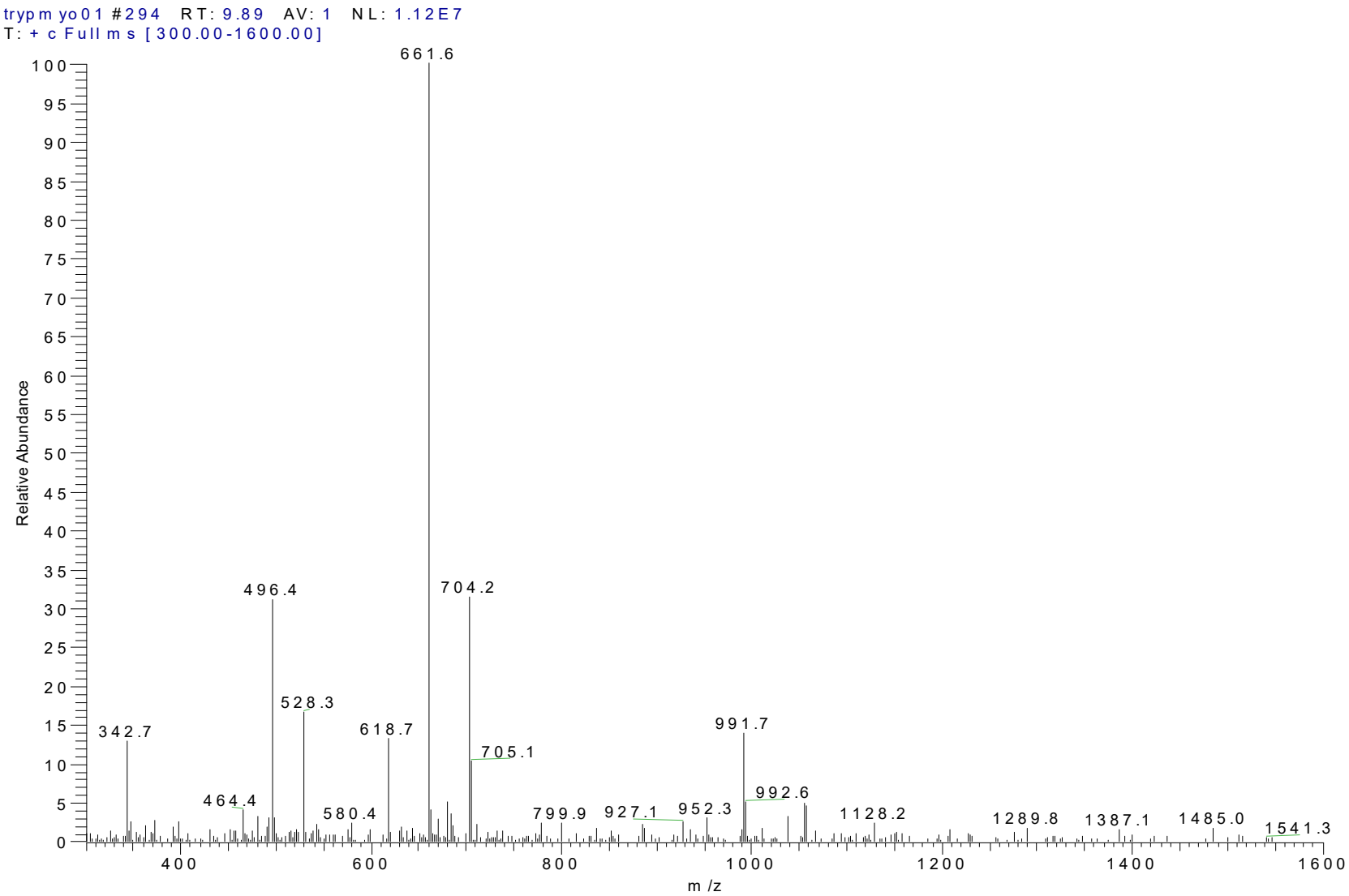
Ion Trap

Orbitrap

FT-ICR = **F**ourier **T**ransform **I**on **C**yclotron **R**esonance

Electron Multiplier

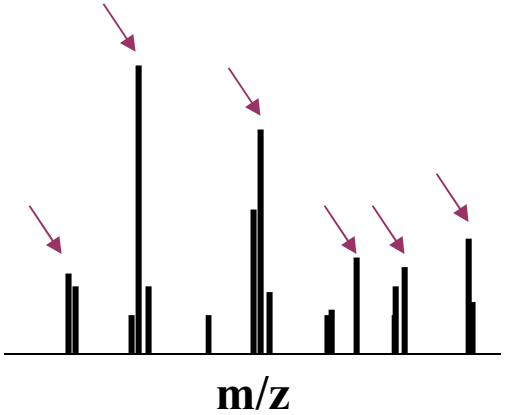
MS Spectrum



Mass spectrometry

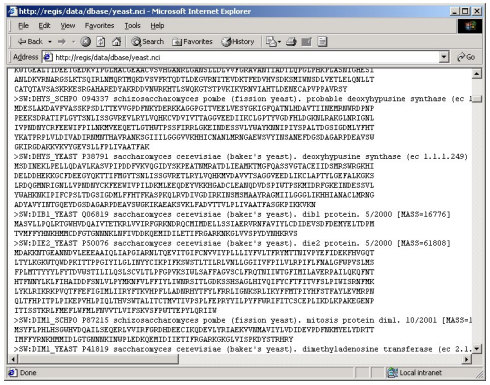
Identifying peptides using an MS spectrum:

List of peptide masses
from MS scan



Search
algorithm

Sequence database

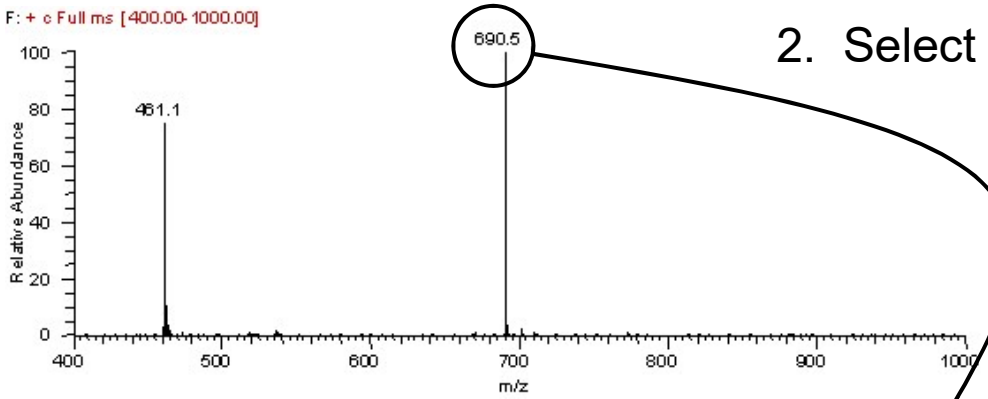


Identified peptide/protein

- Peptide spectrum assignment with Peptide Mass Fingerprinting is only advisable with samples of low complexity and small sequence databases, as the number of all possible peptides with a given mass over charge is huge in large sequence databases.

- Obtaining **sequence information** for a peptide ion:

1. Acquire full (MS) scan

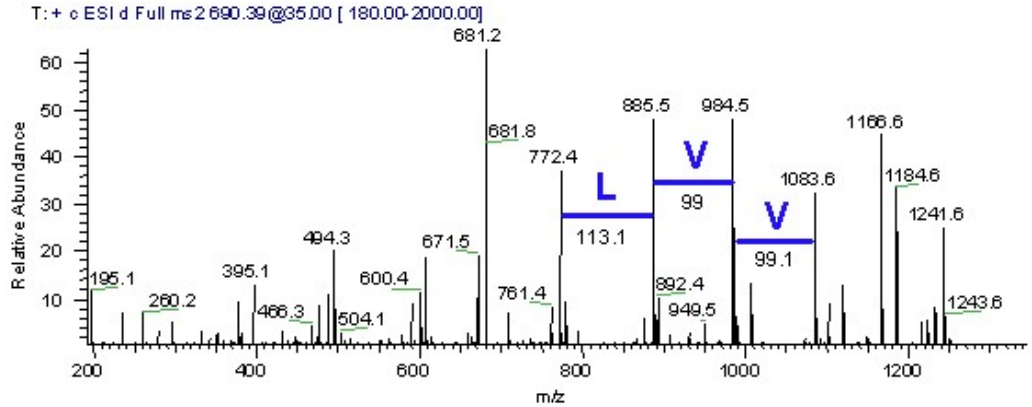


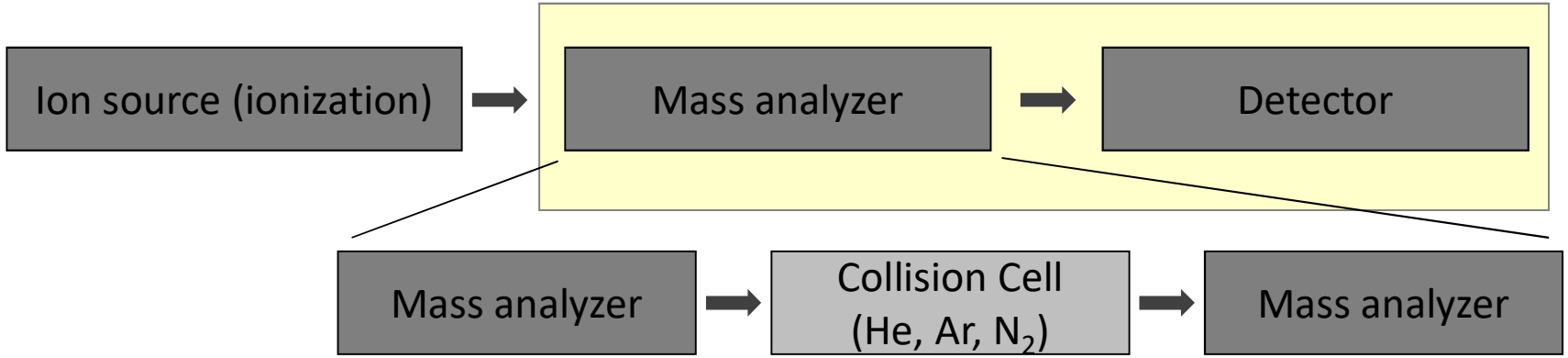
2. Select an ion

3. Isolate the ion

4. Fragment the ion with Collision Induced Dissociation (CID)

5. Acquire MS/MS scan

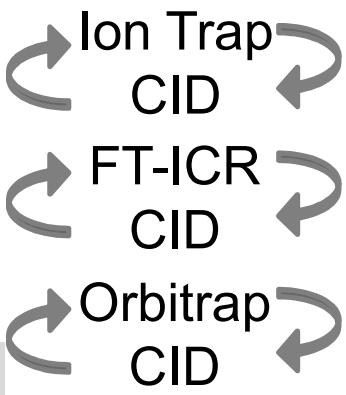




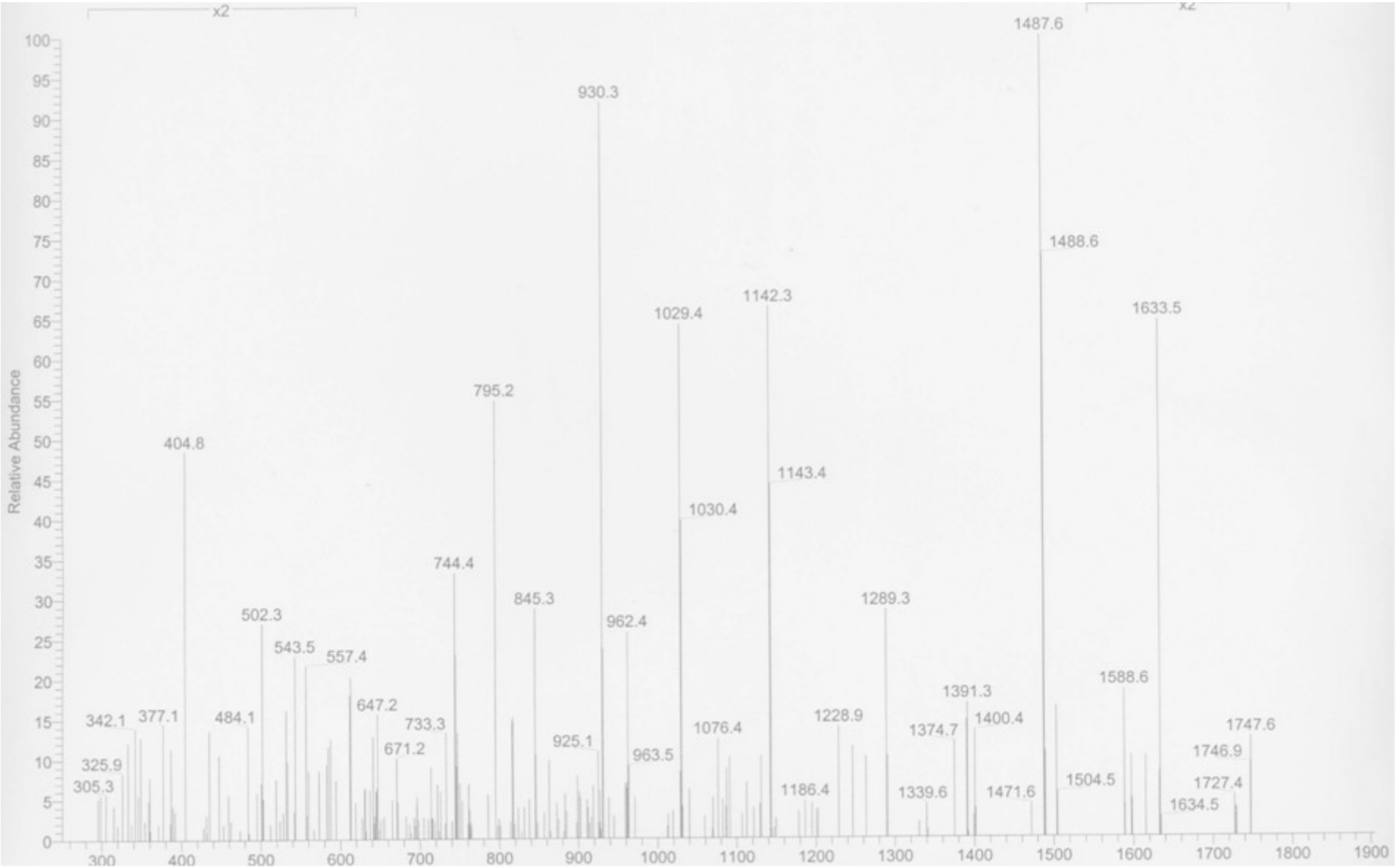
MS/MS in space

TSQ:	Quadrupole	➡	CID	➡	Quadrupole
Q-TOF:	Quadrupole	➡	CID	➡	TOF
TOF/TOF:	TOF	➡	CID	➡	TOF

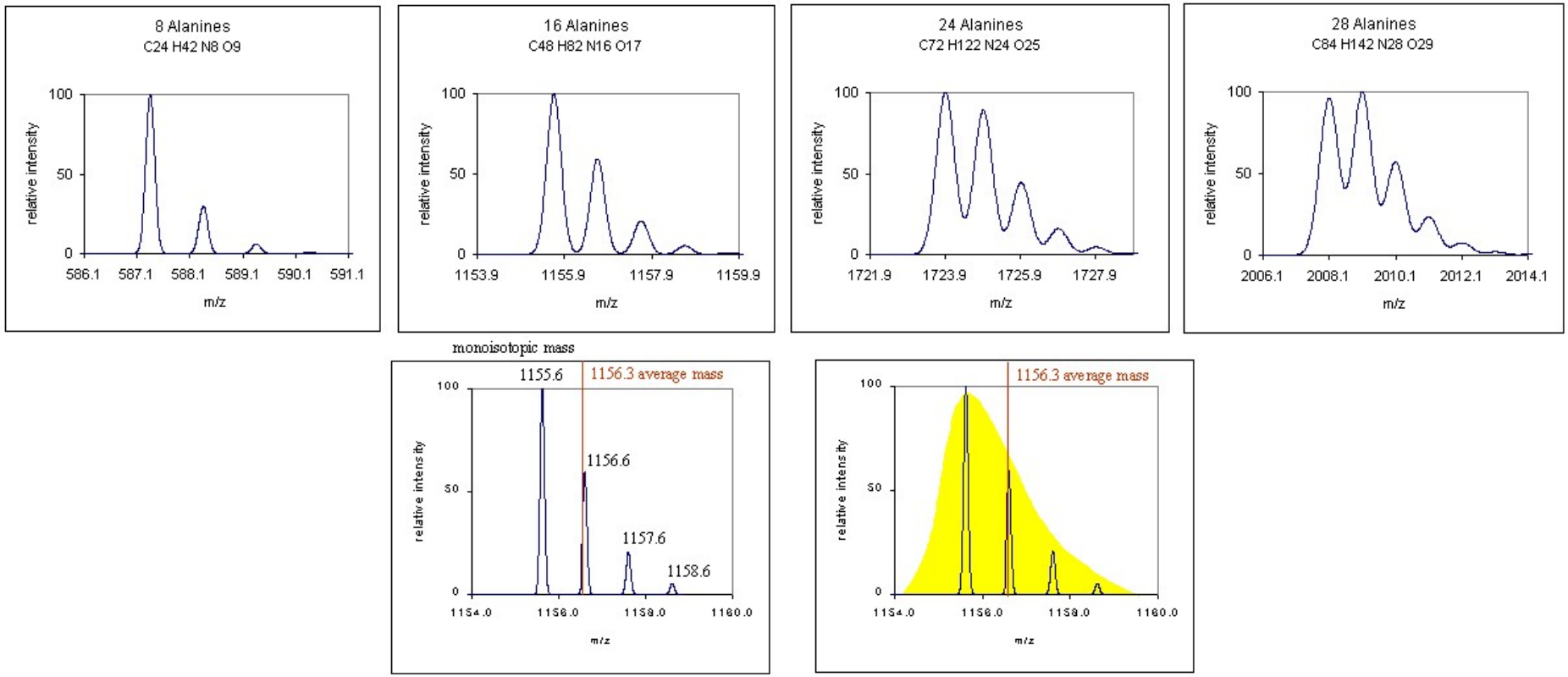
MS/MS in time



MS/MS spectrum

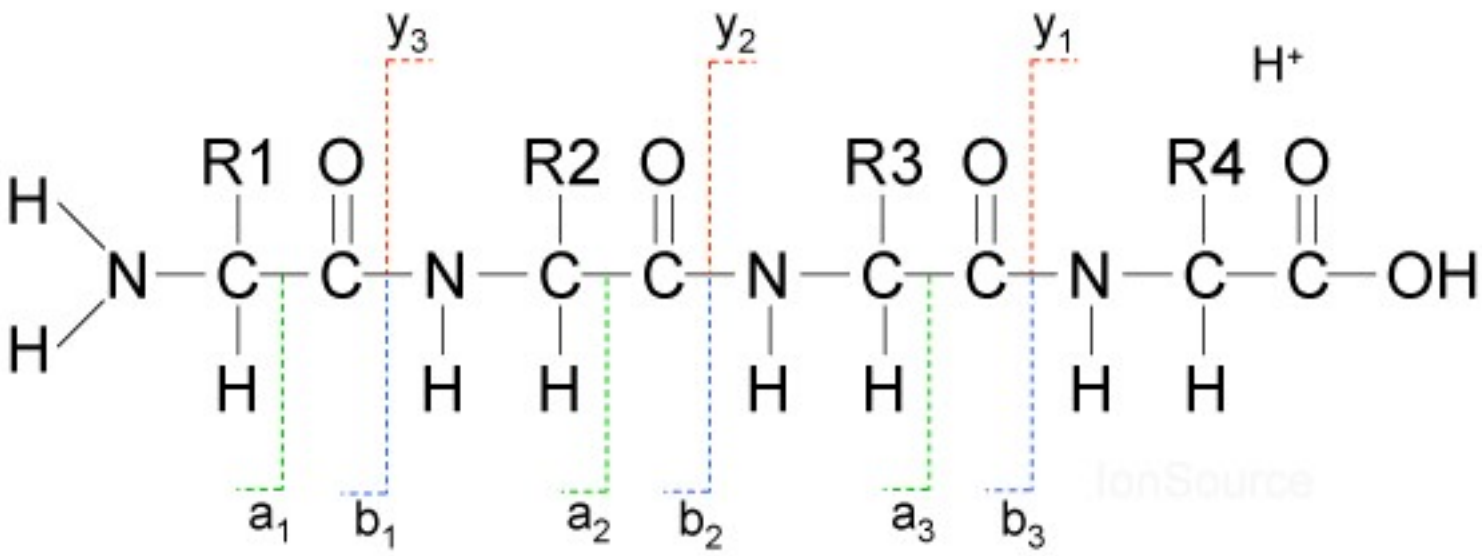


The isotope issue



- 1/100 C atoms is C13
- The more atoms a peptide contains, the more probably it is that one to several C atoms are C13

a, b and y ions after CID



- 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 (N-terminus), and **y** ions appear to extend from the carboxyl terminus (C-terminus).
- **a** ions occur at a lower frequency and abundance in relation to **b** ions.

adjusted from http://www.ionsource.com/

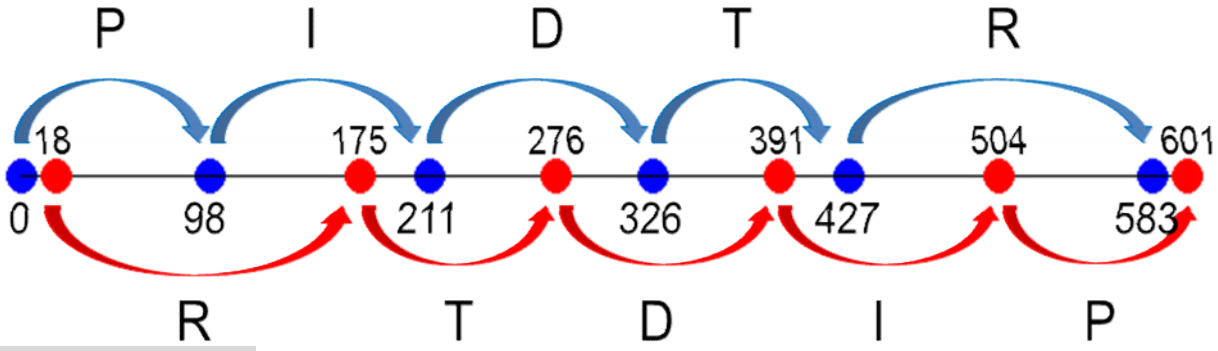
P – I – D – T – R

m/z = 601.31

Masses of b- and y- ions:

b-ions				y-ions
		PIDTR	601.31	y5
b1	98.05	P - - - - - IDTR	504.26	y4
b2	211.14	PI - - - - - DTR	391.18	y3
b3	326.16	PID - - - - - TR	276.15	y2
b4	427.21	PIDT - - - - - R	175.10	y1
b5	583.31	PIDTR		

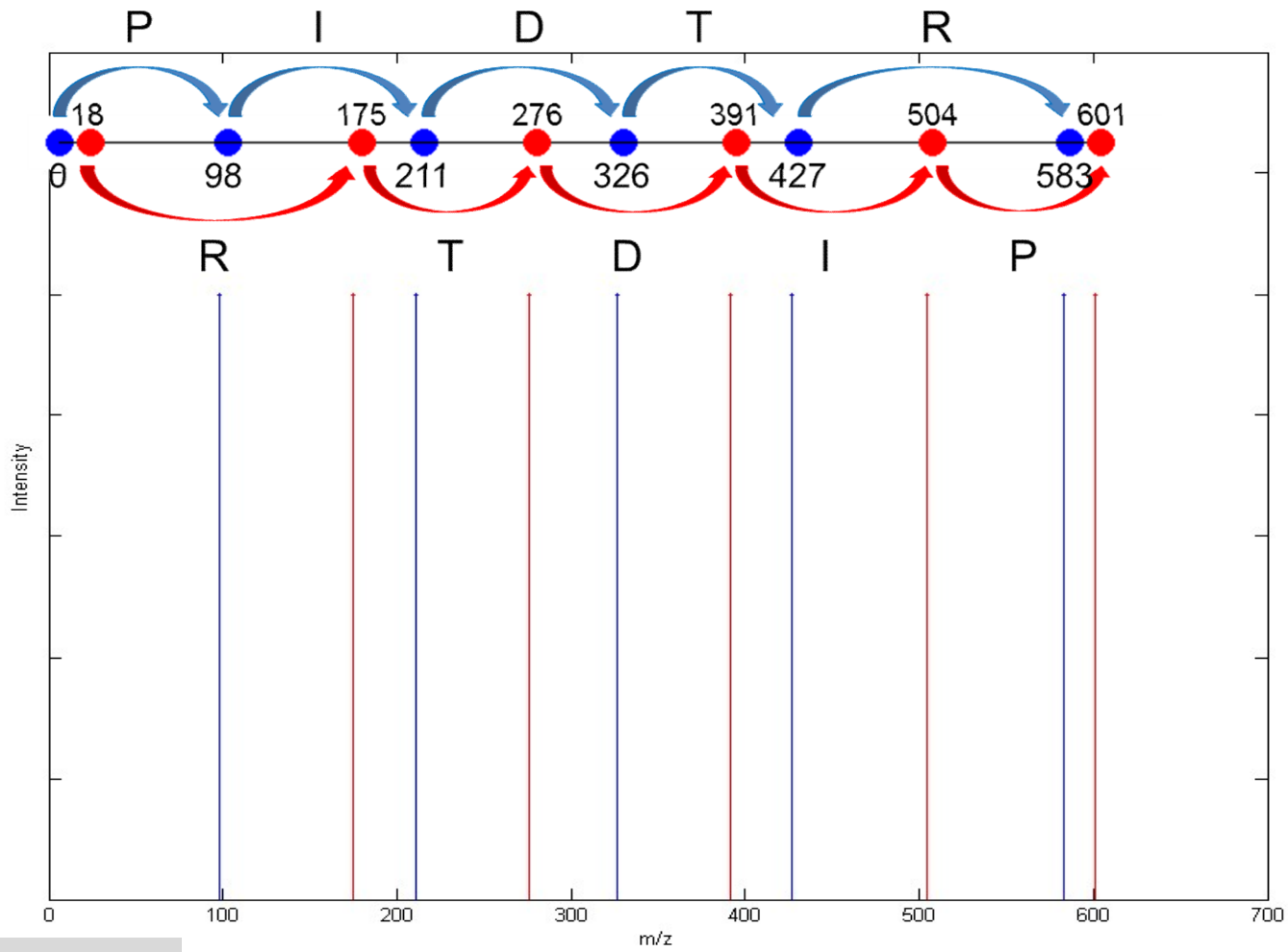
Fragment masses aligned along a spectrum graph:



Peptide spectrum assignment

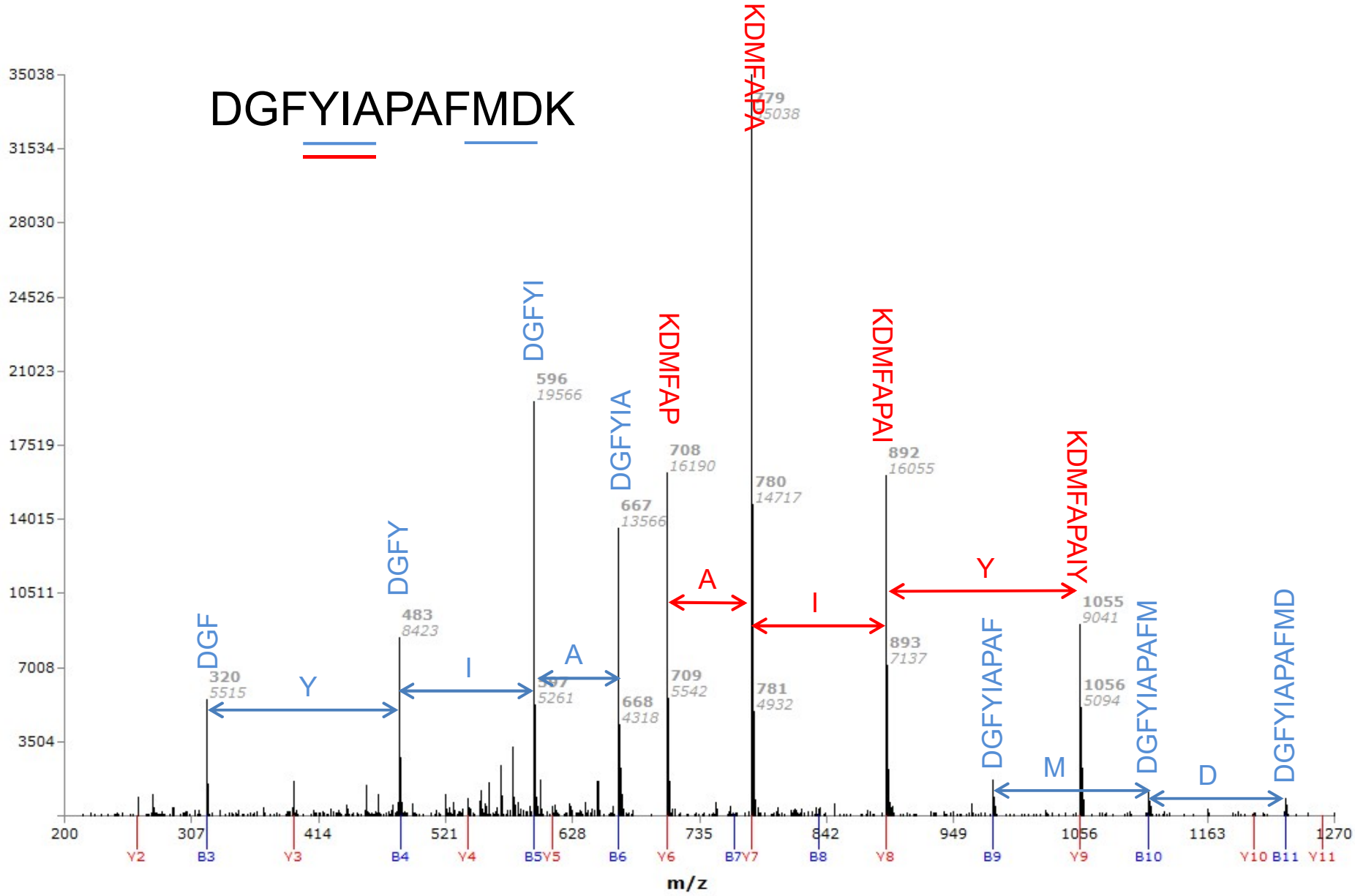
P – I – D – T – R

m/z = 601.31



Peptide spectrum assignment

Analysing a spectrum for the fragments



De novo peptide sequencing

- All information about the peptide sequence resides in the MS/MS spectrum
- Database independent
- Can either be done manually or with algorithms

Manual *de novo* sequencing

- M = Mass of the peptide
- **Precursor ion** $M_{Precursor} \cong \frac{(M+2H)^{2+}}{2}$
- **Parent ion** $M_{Parent} \cong (M+H)^{+}$

$$M_{Parent} = M_{Precursor} * 2 - 1 = \frac{(M+2H)^{2+}}{2} * 2 - 1$$

- Monoisotopic mass of the parent ion

$$M_{Parent\ mono} = M_{Parent\ average} - \frac{M_{Parent\ average}}{1463}$$

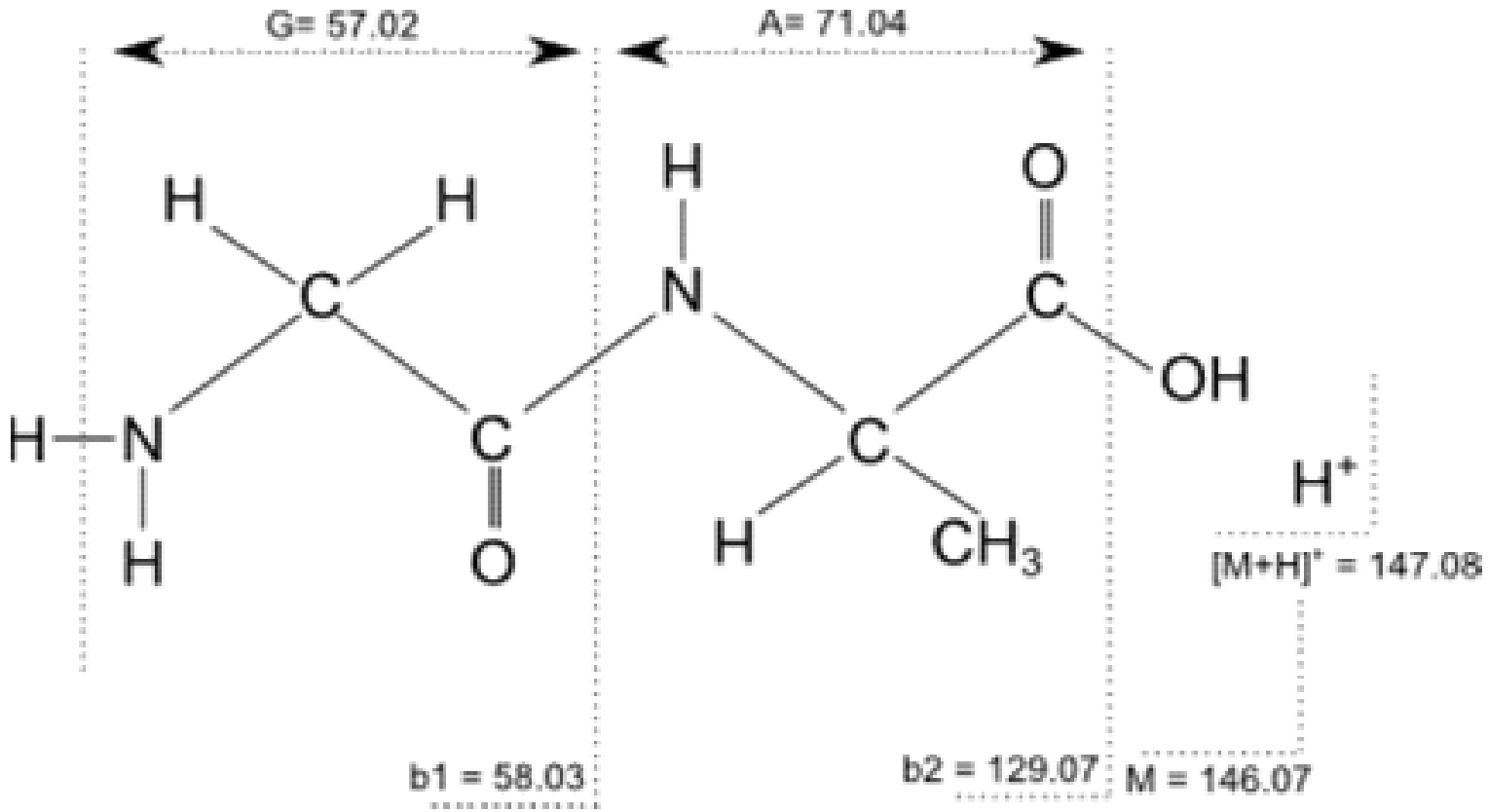
Zubarev and Bondarenko, 1991

- b- and y- ions:

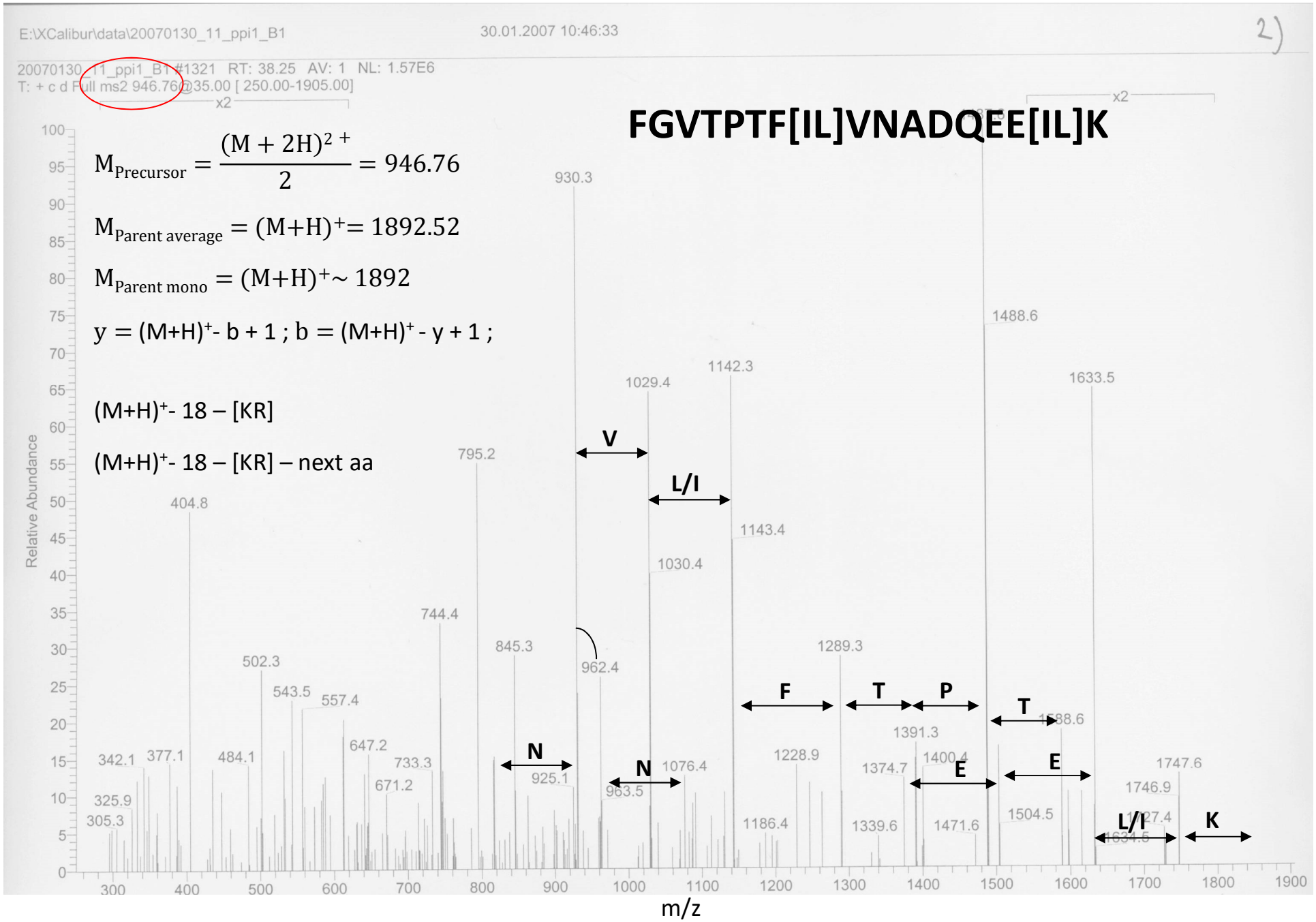
$$y = (M+H)^{+} - b + 1 \ ; \ b = (M+H)^{+} - y + 1$$

AA Codes		Mono.	AA Codes		Mono.
Gly	G	57.021464	Asp	D	115.02694
Ala	A	71.037114	Gln	Q	128.05858
Ser	S	87.032029	Lys	K	128.09496
Pro	P	97.052764	Glu	E	129.04259
Val	V	99.068414	Met	M	131.04048
Thr	T	101.04768	His	H	137.05891
Cys	C	103.00919	Phe	F	147.06841
Leu	L	113.08406	Arg	R	156.10111
Ile	I	113.08406	CMC		161.01467
Asn	N	114.04293	Tyr	Y	163.06333
IonSource sm mass spectrometry educational resource			Trp	W	186.07931

- Find the b-ion without the C-terminal K or R: $(M+H)^{+}$ - 18 ('lost' oxygen) – [KR]



- Upon peptide bond formation, H_2O gets released and this mass is therefore not included in the Δ mass for the individual amino acids
- The largest b-ion is mass of the parent ion – 18, because H_2O gets released and the positive charge resides at the C-terminus


E:\XCalibur\data\20070130_11_ppi1_B1
30.01.2007 10:46:33
2)
20070130_11_ppi1_B1 #1321 RT: 38.25 AV: 1 NL: 1.57E6
T: + c d Full ms2 946.76@35.00 [250.00-1905.00]
x2
FGVTPTF[IL]VNADQEE[IL]K
$M_{Precursor} = \frac{(M + 2H)^{2+}}{2} = 946.76$
$M_{Parent\ average} = (M+H)^{+} = 1892.52$
$M_{Parent\ mono} = (M+H)^{+} \sim 1892$
$y = (M+H)^{+} - b + 1$; $b = (M+H)^{+} - y + 1$;
$(M+H)^{+} - 18 - [KR]$
$(M+H)^{+} - 18 - [KR]$ – next aa
Relative Abundance
m/z
V
L/I
N
N
F
T
P
T
E
E
L/I
K
930.3
1488.6
1142.3
1029.4
1633.5
795.2
404.8
1143.4
1030.4
744.4
845.3
962.4
1289.3
502.3
543.5
557.4
1391.3
1588.6
1228.9
1400.4
1374.7
1747.6
342.1
377.1
484.1
647.2
733.3
1076.4
925.1
963.5
671.2
1746.9
325.9
305.3
1186.4
1339.6
1471.6
1504.5

Peptide fragmentation

ion	offset[b]	Δ[c]	no. peaks[d]	no. spectra[e]	probability
y	19.020	0.002	2245/2792	376	0.804
b	1.006	−0.002	1934/2806	374	0.689
$b-H_2O$	−17.005	−0.002	777/2744	264	0.283
$y/2^f$	9.508	−0.001	508/2359	293	0.215
$y-H_2O$	1.005	−0.003	312/2360	211	0.132
y^{+2}	10.012	−0.001	316/2448	215	0.129
$b-NH_3$	−16.021	−0.002	253/2746	119	0.092
a	−26.988	−0.001	205/2706	144	0.076
$[y-H_2O]^{+2}$	1.006	−0.002	156/2246	127	0.070
$[y-H_2O-H_2O]^{+2}$	−7.998	0.000	142/2189	134	0.065
$b-H_2O-H_2O$	−35.015	−0.002	119/2661	60	0.045
$y-NH_3$	1.989	−0.003	110/2689	79	0.041
$[y-H_2O-NH_3]^{+2}$	−7.507	−0.001	75/2192	73	0.034
$b/2^f$	0.503	−0.001	64/2139	42	0.030
$b-H_2O-NH_3$	−34.031	−0.002	71/2663	42	0.027
$a-NH_3$	−44.015	−0.002	42/2652	38	0.016
$a-H_2O$	−44.999	−0.001	32/2650	25	0.012
$[y-NH_3]^{+2}$	1.498	−0.001	23/2248	20	0.010
b^{+2}	1.006	−0.002	14/2146	12	0.007
$b-NH_3-NH_3$	−33.047	−0.002	17/2664	11	0.006
$y-H_2O-H_2O$	−17.007	−0.004	12/2673	11	0.005
$y-H_2O-NH_3$	−16.022	−0.003	10/2676	10	0.004
Internal+H	1.005	−0.003	227/10841	144	0.021
Internal+$H-H_2O$	−17.005	−0.002	125/10345	84	0.012
Internal+NH_2+H_2O	34.027	0.002	112/11633	92	0.010

Frank *et al.*, 2007

- Peptides do not fragment sequentially, but fragmentation events are somewhat random
- Some fragmentations are preferred over others as noted by the variation in the abundance of observed peaks
- The mess of peaks normally observed in a fragment spectrum are a reflection of the population of fragment ions produced in the collision cell of a mass spectrometer

Neutral losses and peptide modifications

- Neutral losses
- H_20 (-18 Da)
- NH_2 (-17 Da)
- CO (-28 Da)
- H_3PO_4 (-98 Da)

- Modifications

Oxidation: M +16 Da

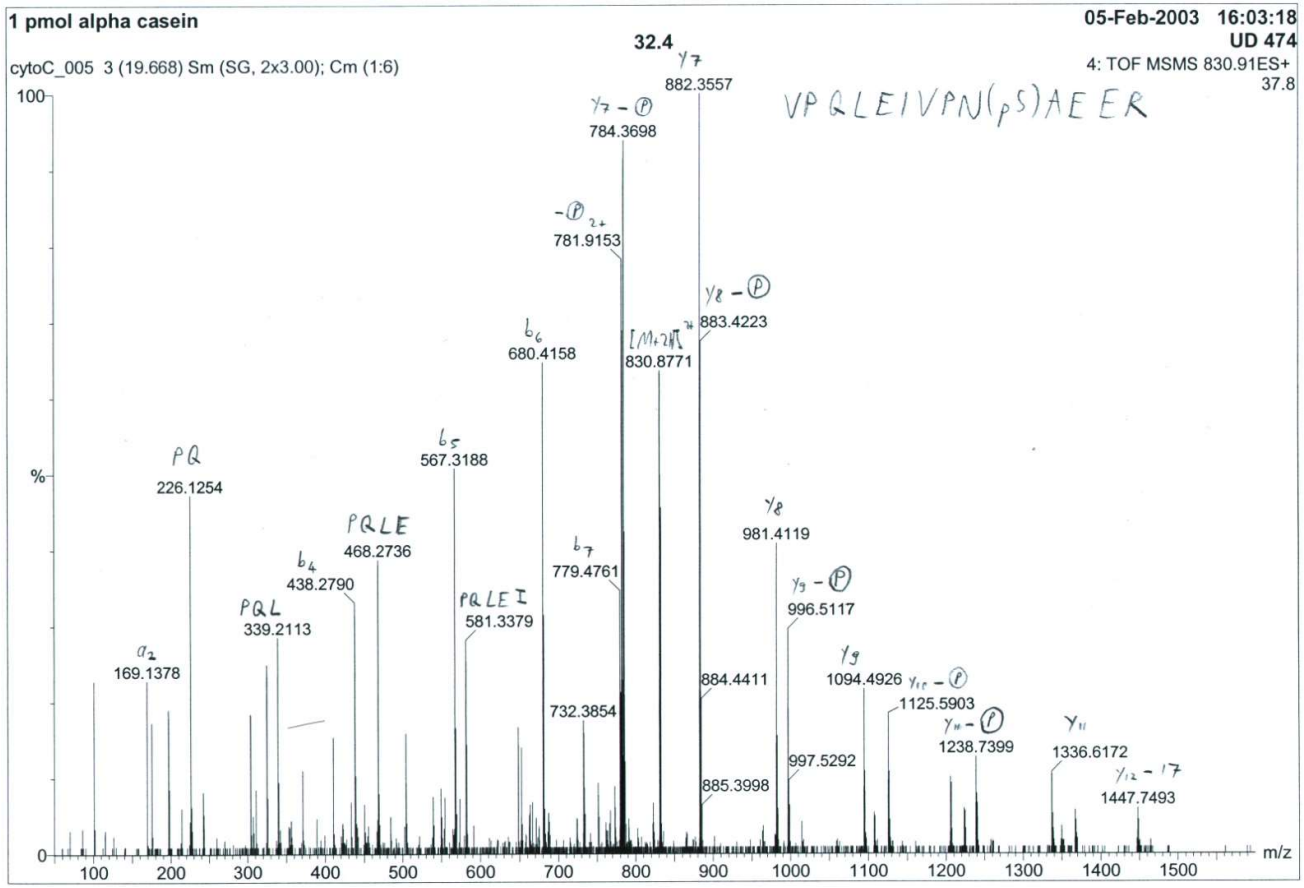
Deamidation: N → Q, -1 Da

Methylation: R +14 Da

Phosphorylation: S, T, Y +80 Da

Acetylation: K, R, N-terminus +42 Da

Carboxyamidomethylation: C +57 Da



Peptide spectrum assignment

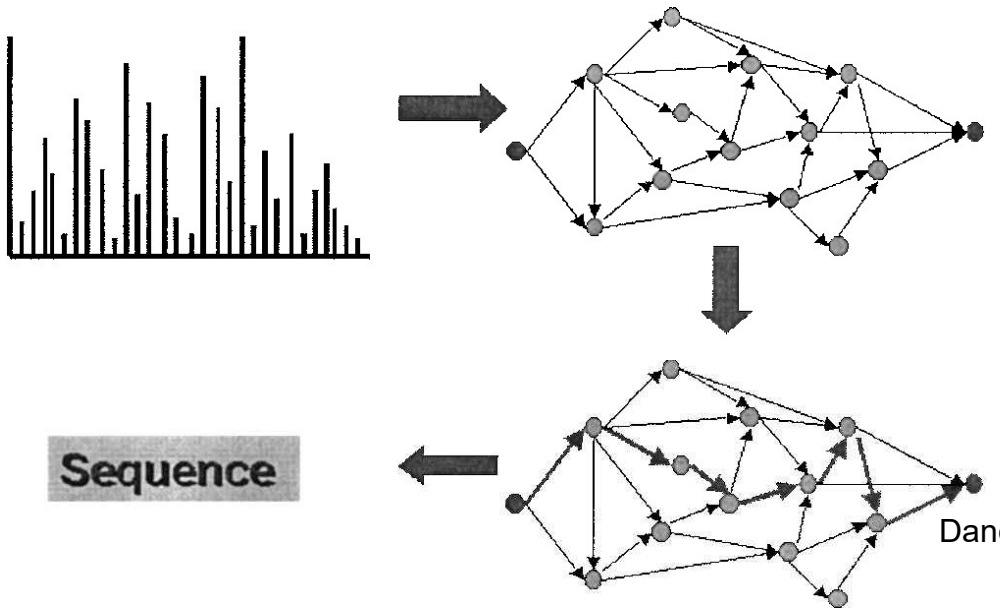
Dynamic programming in *de novo* sequencing

- In de novo sequencing with dynamic programming, the peaks in a spectrum are transformed to a spectrum graph representation. In the spectrum graph representation, the peaks in the spectrum serve as nodes in the graph, while the edges of the graph link nodes differing by the mass of an amino acid.

- Each peak in an experimental spectrum is transformed into several nodes in a spectrum graph, and each node represents a possible fragment type assignment for a peak

- Sequence reconstructions correspond to paths in the spectrum graphs

- In correct reconstructions the nodes in the path correspond to cleavages in the peptide



Dancik *et al.*, 1999

Peptide spectrum assignment

Use of Hidden Markov Models in *de novo* sequencing

Anal. Chem. **2005**, *77*, 7265–7273

NovoHMM: A Hidden Markov Model for de Novo Peptide Sequencing

Bernd Fischer,[†] Volker Roth,[†] Franz Roos,[§] Jonas Grossmann,[‡] Sacha Baginsky,[‡] Peter Widmayer,[§] Wilhelm Gruissem,[‡] and Joachim M. Buhmann[†]

- In the NovoHMM model, the observable random variables correspond to the observed mass peaks, whereas the hidden variables represent the unknown underlying sequence

1. Dynamic programming
   - The algorithms suffer from ‘real life issues’ of peptide mass spectrometry, e.g. they are sensitive against noisy data
   - Requires pre-processing of the information in an MS/MS spectrum
2. Hidden Markov Models
   - Fully probabilistic
   - Can deal with noisy data

→ Studies comparing the performance of *de novo* sequencing algorithms revealed that the rate of exact peptide sequence identification is low with high error rate

→ Problems for *de novo* sequencing are limited mass accuracy of the mass spectrometers, missing ions, unknown identity of the peaks and additional, sequence-independent peaks

Database search programs

Different database-dependent peptide identification search algorithms:

- Sequest
- Mascot
- PepSplice
- OMSSA
- X!Tandem
- Phenyx
- ProteinPilot
- SpectrumMill
- ProbID
- PepFrag
- InSpect
- ...

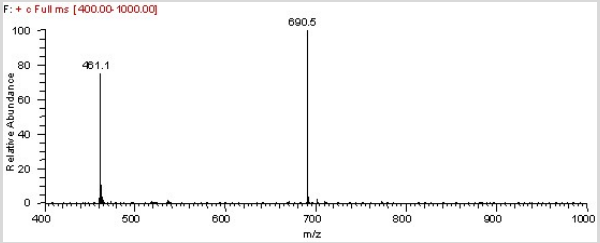
- Peptide sequences with one or more scores with which to evaluate the likelihood that the resulting sequence is correct.
- Even though each implementation is different, they operate under the same general principle.

Database-dependent peptide identification

Experimental

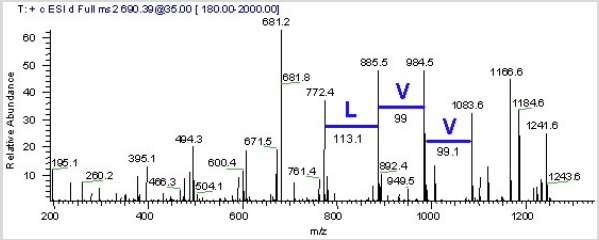
Protein sample

tryptic digest



Full Scan [MS]:
Mass measurement of full peptides

fragmen-tation



MS/MS-Scan:
Measurement of peptide fragments

1) Peptides from a protein database are matched to the measured mass

2) Theoretical spectrum is cross-correlated to the measured one

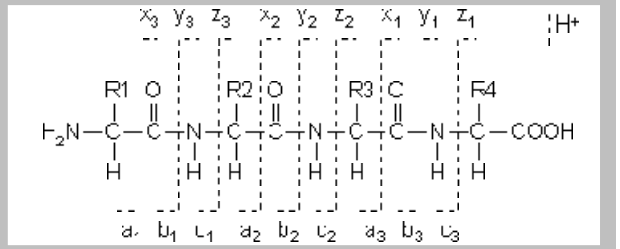
Protein database

in silico digest

Peptide database

Theoretical tryptic peptides
from a protein database

in silico dissociation



Theorectical spectra
assuming peptide bond breakage

In silico

Peptide spectrum assignment

- The goal is to identify the best sequence match to the spectrum
- The details of this implementation differ among the algorithms. In addition, the methods used to assign scores are very different.
- Four basic approaches have been developed to model matches to the sequences: descriptive models, interpretative models, stochastic models and statistical and probability models (Sadygov, Cociorva and Yates, 2004)

Descriptive algorithms are based on a mechanistic prediction of how peptides fragment in a tandem mass spectrometer, which is then quantified to determine the quality of the match between the prediction and the experimental spectrum. Mathematical methods such as correlation analysis have been used to assess match quality.

- Sequest is an example of a program using a descriptive model:
  - S_p, sums the peak intensity of fragment ions matching the predicted sequence ions and accounts for the continuity of an ion series and the length of a peptide
  - Xcorr, is a cross-correlation score of the experimental and theoretical spectra
  - ΔCn gives the normalised difference of Xcorr values between the best sequence and lower-scoring matches and is useful to determine the uniqueness of the match

Sadygov, Cociorva and Yates, 2004

Interpretative approaches are based on manual or automated interpretation of a partial sequence from a tandem mass spectrum and incorporation of that sequence into a database search. Matches between the sequence and the spectrum have been scored using probabilities or correlation methods.

- PeptideSearch and InSpect belong to the programs using sequence tagging:
  - The program identifies a continuous series of fragment ions (sequence tag)
  - Every candidate peptide is divided into three parts: m1, sequence tag = m2, m3
  - The sequence tag can be from the b- or y-ion series and therefore both possibilities have to be considered by the algorithm
  - The algorithm searches the database for matches using the masses of m1, m2 and m3, as well as information from protease specificity
  - The sequence match is then scored by calculating the probability of the match being non-random

Sadygov, Cociorva and Yates, 2004

Stochastic models are based on probability models for the generation of tandem mass spectra and the fragmentation of peptides. Basic probabilities of fragment ion matches are obtained from training sets of spectra of known sequence identity. Stochastic models use statistical limits on the measurement and fragmentation process to create a likelihood that the match is correct.

- SCOPE is an example of a program using a stochastic model:
  - First, the fragmentation probabilities are estimated with assumptions on fragmentation patterns and/or with collections of annotated spectra
    - probability of obtaining the fragmentation pattern F from CID of peptide p
  - Second, the probability of observing a collection of spectral peaks given a particular peptide fragmentation is computed
    - probability of fragmentation pattern F to generate spectrum S
  - Finally, the probability of obtaining spectrum S from peptide b is computed

Sadygov, Cociorva and Yates, 2004

Statistical and probability models determine the relationship between the tandem mass spectrum and sequences. The probability of peptide identification and its significance are then derived from the model.

- A hypergeometric probability models the frequencies of database peptides based on the number of matches. The significance of a peptide match is defined as a rejection of the null hypothesis that all fragment matches are random

- In Mascot, the score is the probability-based MOWSE (MOlecular Weight SEarch) score given as s S = - 10*log(P) where P is the probability that the observed match between experimental data and a peptide sequence is a random event

- The correct match, which is a not a random event, has a very low probability

Sadygov, Cociorva and Yates, 2004

Expect value = the number of times you could expect to get this score or better by chance

- A completely random match has a score of 1 or higher
- The better the match the lower the expect value

Empirical statistical model to estimate the accuracy of peptide identifications

HTML-SUMMARY v.8 (rev 0), Copyright 1996 - Mozilla Firefox

file://///Snapserver/proteomesoft/users/brian/sergei_digest_A_full_01.html

HTML-SUMMARY v.8 (rev 0), Copyright 1996
Molecular Biotechnology, Univ. of Washington, J.Eng/J.Yates
Compiled for use by the Aebersold lab @ Univ. of Washington
03/19/02, 07:20 AM, /data/search/akeller/databases/new_hum_plus_proteinmix.db, AVG/AVG

#	File	MH+	XCorr	dCn	Sp	RSp	Ions	Ref	Sequence
325	./sergei_digest_A_full_01.1001.1003.3	2511.7 (-0.6)	7.5863	0.450	4546.1	✓ 1	42/ 84	sp|P00921|CAH2_BOVIN	R.MVNNGHSFNVEYDDSQDKAVLK.D
462	./sergei_digest_A_full_01.1239.1241.3	2584.7 (+2.0)	6.1808	0.382	3126.2	✓ 1	39/ 84	sp|P00921|CAH2_BOVIN	R.LVQFHFHWGSSBBQGSEHTVDR.K
1070	./sergei_digest_A_full_01.2335.2337.3	2254.5 (+0.5)	6.0682	0.491	2166.9	✓ 1	37/ 84	sp|P00921|CAH2_BOVIN	K.YGDFGTAAQQPDGLAVVGVFLK.V
1105	./sergei_digest_A_full_01.2405.2407.3	2254.5 (+0.8)	6.0041	0.511	1873.6	✓ 1	35/ 84	sp|P00921|CAH2_BOVIN	K.YGDFGTAAQQPDGLAVVGVFLK.V
510	./sergei_digest_A_full_01.1317.1325.3	2584.7 (+1.5)	5.9521	0.403	2488.3	✓ 1	38/ 84	sp|P00921|CAH2_BOVIN	R.LVQFHFHWGSSBBQGSEHTVDR.K
1219	./sergei_digest_A_full_01.2617.2619.3	2187.6 (-0.1)	5.7343	0.502	2282.1	✓ 1	33/ 72	sp|P02666|CASB_BOVIN	R.DMPIQAFLLYQEPVLGPVR.G
894	./sergei_digest_A_full_01.2013.2015.3	2314.7 (+0.0)	5.5636	0.418	1260.3	✓ 1	33/ 76	sp|P02754|LACB_BOVIN	R.VYVEELKPTPEGDLEILLQK.W
812	./sergei_digest_A_full_01.1873.1875.3	2314.7 (+0.5)	5.5466	0.428	1407.6	✓ 1	35/ 76	sp|P02754|LACB_BOVIN	R.VYVEELKPTPEGDLEILLQK.W
1142	./sergei_digest_A_full_01.2471.2473.2	2254.5 (-0.8)	5.5372	0.526	771.0	✓ 1	24/ 42	sp|P00921|CAH2_BOVIN	K.YGDFGTAAQQPDGLAVVGVFLK.V
856	./sergei_digest_A_full_01.1943.1945.3	2314.7 (+0.9)	5.4579	0.426	1581.3	✓ 1	34/ 76	sp|P02754|LACB_BOVIN	R.VYVEELKPTPEGDLEILLQK.W
1289	./sergei_digest_A_full_01.2765.2771.2	2709.1 (+0.4)	5.3678	0.495	1654.1	✓ 1	24/ 50	sp|P02754|LACB_BOVIN	K.VAGTWYSLAMAASDISLLDAQSAPLR.V
1220	./sergei_digest_A_full_01.2621.2623.2	2187.6 (-0.5)	5.3391	0.461	1646.5	✓ 1	22/ 36	sp|P02666|CASB_BOVIN	R.DMPIQAFLLYQEPVLGPVR.G
1153	./sergei_digest_A_full_01.2491.2493.3	2219.5 (+2.1)	5.3167	0.276	1640.4	✓ 1	31/ 72	sp|P00921|CAH2_BOVIN	R.TLNFNAEGEPELLMLANWR.P
1102	./sergei_digest_A_full_01.2401.2403.2	2254.5 (-0.6)	5.1675	0.495	1009.0	✓ 1	24/ 42	sp|P00921|CAH2_BOVIN	K.YGDFGTAAQQPDGLAVVGVFLK.V
1067	./sergei_digest_A_full_01.2329.2333.2	2254.5 (-0.9)	5.1492	0.546	688.2	✓ 1	23/ 42	sp|P00921|CAH2_BOVIN	K.YGDFGTAAQQPDGLAVVGVFLK.V
125	./sergei_digest_A_full_01.0681.0681.2	2100.2 (+0.8)	4.9146	0.481	1779.2	✓ 1	22/ 34	sp|P00921|CAH2_BOVIN	R.MVNNGHSFNVEYDDSQDK.A
1020	./sergei_digest_A_full_01.2237.2239.2	1568.7 (-0.1)	4.8921	0.413	1764.0	✓ 1	19/ 24	sp|P02769|ALBU_BOVIN	K.DAFLGSFLYEYSR.R
981	./sergei_digest_A_full_01.2163.2165.2	2147.3 (-0.7)	4.8738	0.452	1515.4	✓ 1	26/ 38	sp|P02666|CASB_BOVIN	E.LNVPGEIVESLSSSEESITR.I
533	./sergei_digest_A_full_01.1361.1367.3	2906.0 (+0.7)	4.8712	0.301	655.0	✓ 1	29/ 92	sp|P02666|CASB_BOVIN	K.FQSEEQQQTEDELQDKIHPFAQTQ.S
815	./sergei_digest_A_full_01.1879.1883.2	2314.7 (-0.6)	4.7827	0.419	438.7	X 1	19/ 38	sp|P02754|LACB_BOVIN	R.VYVEELKPTPEGDLEILLQK.W
760	./sergei_digest_A_full_01.1771.1773.2	2034.2 (-0.1)	4.7587	0.465	1047.9	✓ 1	26/ 36	sp|P02666|CASB_BOVIN	L.NVPGEIVESLSSSEESITR.I
1048	./sergei_digest_A_full_01.2289.2291.2	2804.0 (+0.5)	4.7564	0.433	1497.0	✓ 1	26/ 48	sp|P02666|CASB_BOVIN	A.RELEELNVPGEIVESLSSSEESITR.I
157	./sergei_digest_A_full_01.0735.0737.3	2100.2 (-0.5)	4.7471	0.300	1496.2	✓ 1	32/ 68	sp|P00921|CAH2_BOVIN	R.MVNNGHSFNVEYDDSQDK.A
91	./sergei_digest_A_full_01.0607.0609.2	1983.0 (-0.4)	4.6602	0.358	3224.6	X 1	24/ 30	sp|P02666|CASB_BOVIN	K.FQSEEQQQTEDELQDK.I
625	./sergei_digest_A_full_01.1523.1525.3	2046.3 (-0.7)	4.6539	0.307	1901.4	✓ 1	29/ 60	sp|P02769|ALBU_BOVIN	R.RHPYFYAPELLYYANK.Y
117	./sergei_digest_A_full_01.0663.0665.3	2100.2 (+0.0)	4.5901	0.286	1627.3	X 1	36/ 68	sp|P00921|CAH2_BOVIN	R.MVNNGHSFNVEYDDSQDK.A
712	./sergei_digest_A_full_01.1675.1681.2	1832.0 (-0.3)	4.5772	0.393	1148.8	X 1	21/ 34	sp|Q29443|TRFE_BOVIN	K.GEADAMSLDGGYLYIAGK.C
651	./sergei_digest_A_full_01.1569.1571.3	3106.3 (+0.4)	4.5737	0.362	537.5	X 1	26/100	sp|P02666|CASB_BOVIN	K.FQSEEQQQTEDELQDKIHPFAQTQSL.V
942	./sergei_digest_A_full_01.2085.2089.3	2314.7 (+1.4)	4.5707	0.299	1348.2	X 1	32/ 76	sp|P02754|LACB_BOVIN	R.VYVEELKPTPEGDLEILLQK.W
1045	./sergei_digest_A_full_01.2277.2279.3	2804.0 (-0.5)	4.5012	0.352	1236.7	✓ 1	31/ 96	sp|P02666|CASB_BOVIN	A.RELEELNVPGEIVESLSSSEESITR.I
861	./sergei_digest_A_full_01.1951.1953.2	2314.7 (-0.4)	4.4874	0.348	482.4	X 1	20/ 38	sp|P02754|LACB_BOVIN	R.VYVEELKPTPEGDLEILLQK.W
626	./sergei_digest_A_full_01.1529.1531.2	1480.7 (-0.5)	4.4192	0.315	1731.6	X 1	20/ 24	sp|P02769|ALBU_BOVIN	K.LGEYGFQNALIVR.Y
176	./sergei_digest_A_full_01.0767.0767.2	1983.0 (+0.7)	4.4109	0.414	1966.2	X 1	22/ 30	sp|P02666|CASB_BOVIN	K.FQSEEQQQTEDELQDK.I
328	./sergei_digest_A_full_01.1009.1011.3	1440.7 (+0.8)	4.4102	0.302	2269.3	X 1	28/ 44	sp|P02769|ALBU_BOVIN	R.RHPEYAVSVLLR.L

Done

Peptide spectrum assignment

Empirical statistical model to estimate the accuracy of peptide identifications

Task:

Derive a list of identified peptides from database search results carried out with a large number of MS/MS spectra.

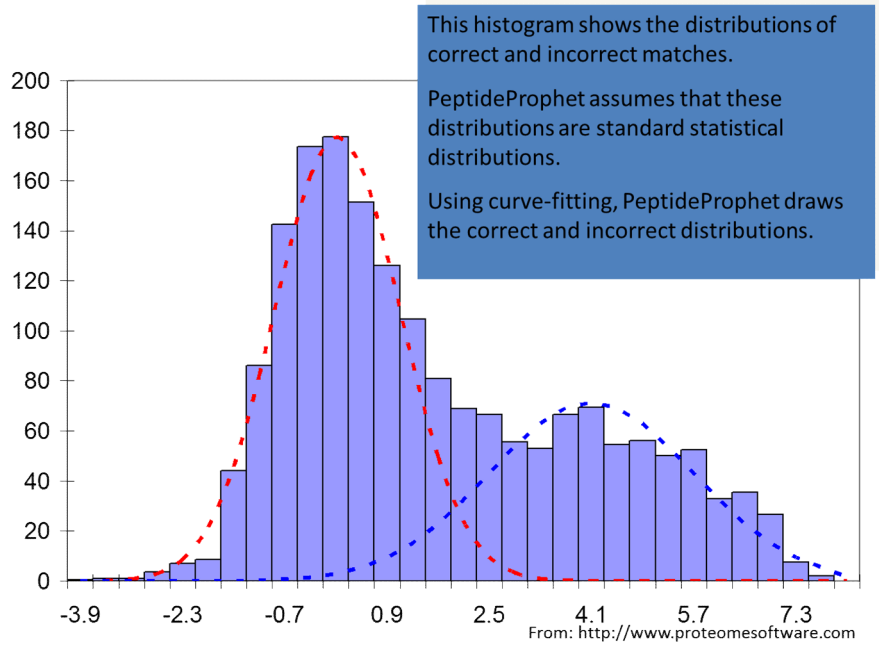
- This entails distinguishing correct peptide assignments from false identifications. For small datasets, this can be achieved by researchers with expertise manually verifying the peptide assignments made by database search programs.

- For high-throughput analysis and consistent data analysis a statistical model is needed to assess the validity of peptide identifications made by MS/MS database searches.

Keller et al., 2002

- **PeptideProphet** computes for each peptide assignment to a spectrum a probability of being correct.

- A discriminant function analysis is used to combine together any number of database search scores into a single discriminant score that best separates training data into correct and incorrect identifications. The discriminant score F is a weighted combination of the database search scores.



- Bayes' theorem gives the probability that a particular peptide assignment with a specific discriminant score is correct:

  $p(+|F) = p(F|+)p(+) / (p(F|+)p(+) + p(F|-)p(-))$

  where $p(+|F)$ = probability that the peptide assignment with discriminant score F is correct

Keller et al., 2002

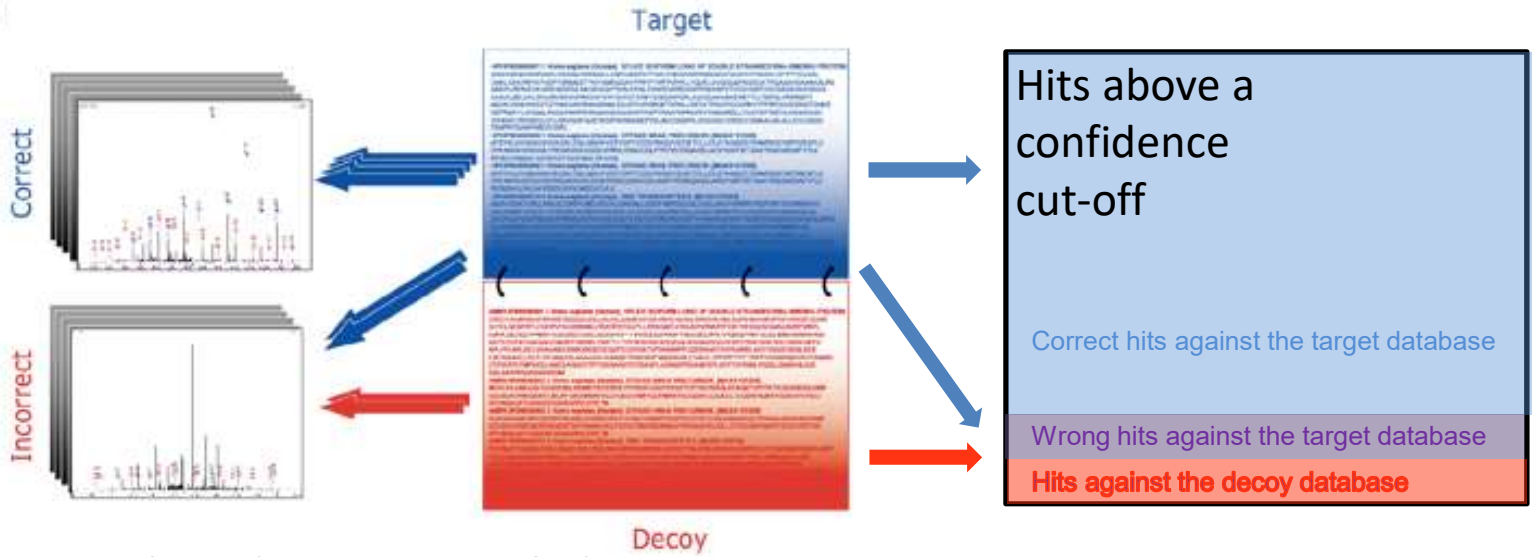
False positives in database searching

False positives are a concern and can occur because:

- Spectra can be single peptide ions, chemical noise, non-peptide molecules for mixtures of co-eluting isobaric peptides
- Peptides are often present at a wide range of concentrations in a sample, and peptides present at the limit of detection can produce poor quality spectra
- Chemistry of peptide fragmentation is not completely understood
- There are amino acid sequences that do not produce a unique fragmentation pattern but share enough of the same fragment ions to be indistinguishable from one another

Target-decoy search strategy

1. All the spectra are searched against a database that consists of the target database concatenated to a decoy database (either randomized or reversed target database)



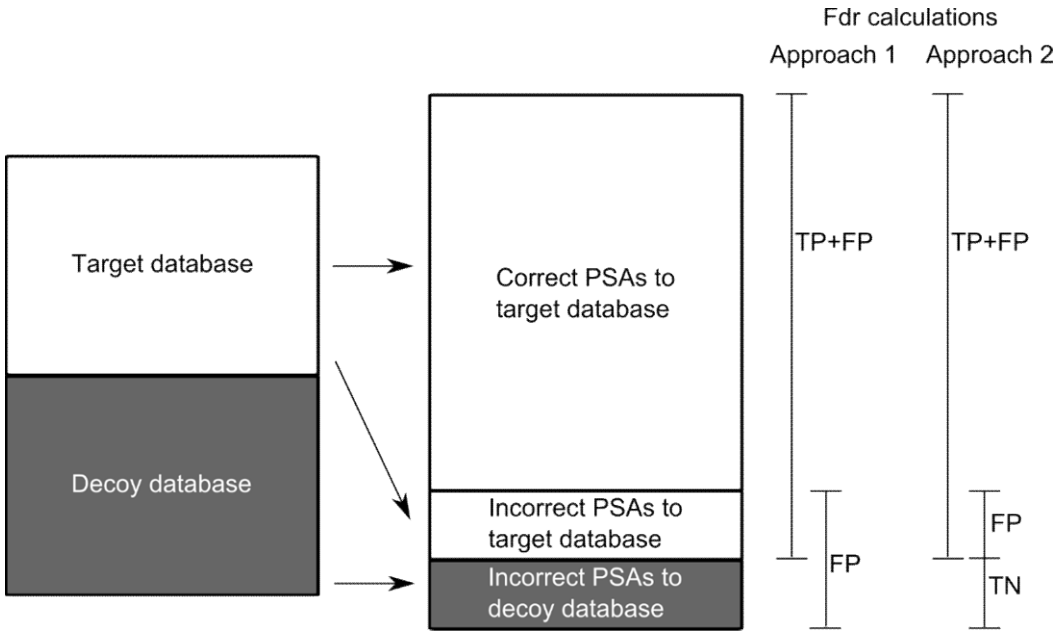
Elias and Gygi, Nature Methods, 2007

1. The hits against the decoy database are clearly wrong as these sequences don't exist

2. It can then be assumed that the number of noticeable wrong hits against the decoy database equals the number of non-noticeable wrong hits against the target database

Calculation of false discovery rates (fdrs)

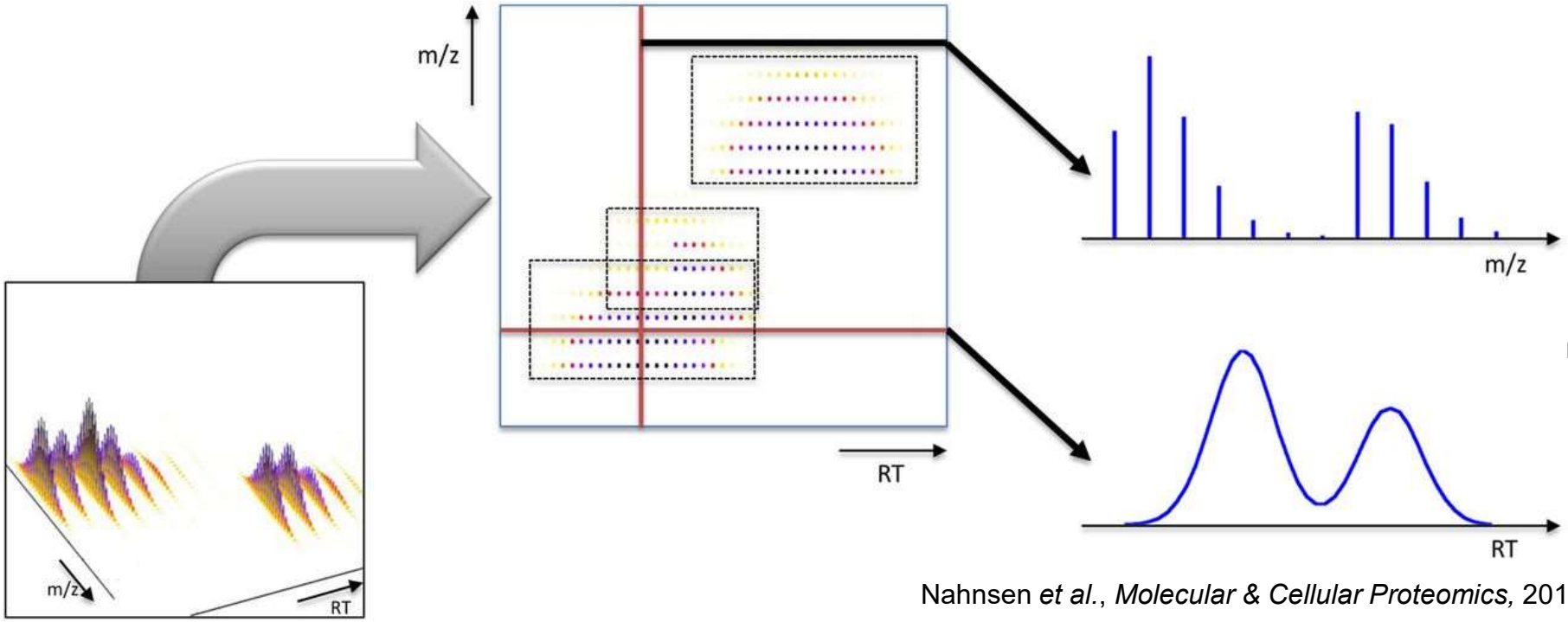
- The estimation of the fdr is a requirement for the analysis and documentation of mass spectrometry data according to the Paris guidelines of Molecular and Cellular Proteomics (Bradshaw, Burlingame, Carr & Aebersold, 2006)



Svozil and Baerenfaller, MIE, 2017

- Global fdrs are calculated for the full dataset
- Local false discovery rates (lfdrs) can be calculated for a subset of the peptide spectrum matches, e.g. the spectra assigned to peptides carrying specific post-translational modifications, or spectra assigned to peptides in an alternative search database, etc.

Label-free quantification (data dependent analysis DDA)

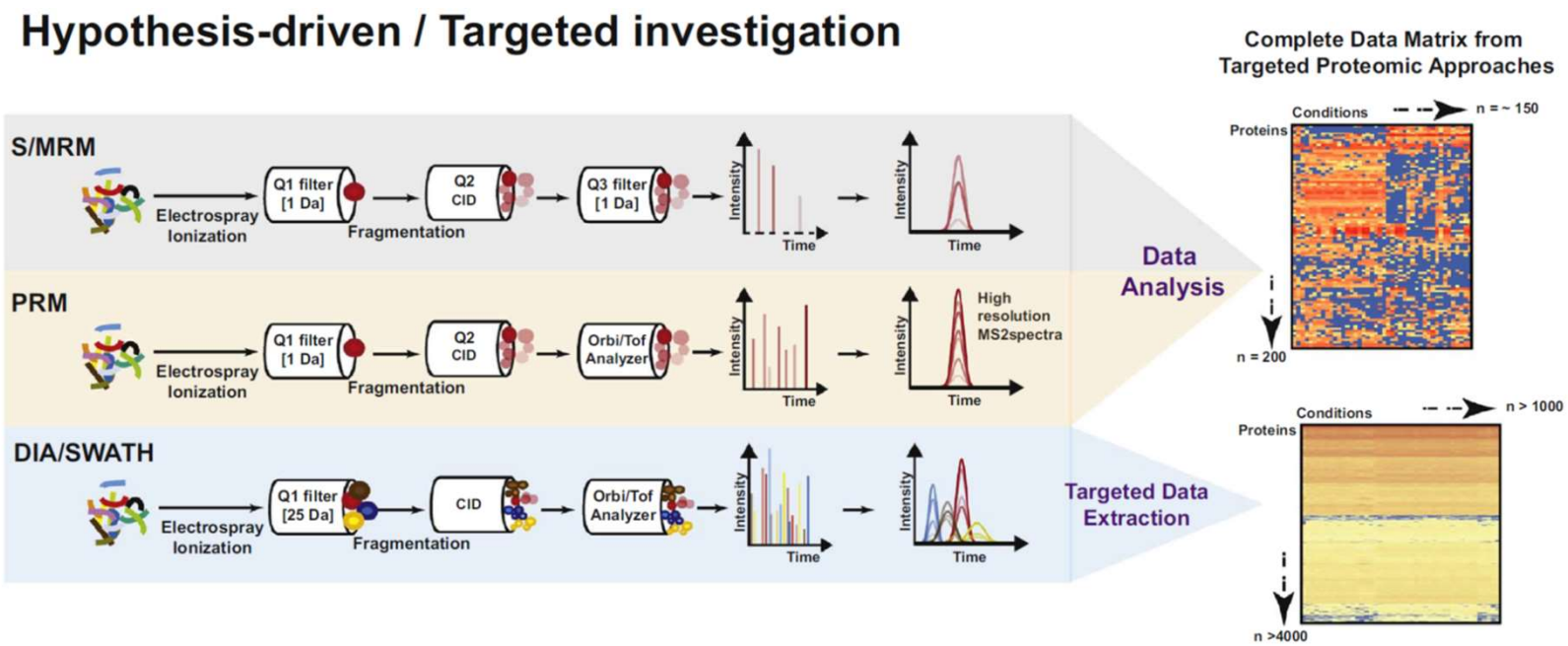


Nahnsen *et al.*, *Molecular & Cellular Proteomics,* 2013

LC/MS data consist of individual MS spectra accumulated over (retention) time. Stacked side by side, these spectra form two-dimensional maps.

→ In spectral counting the basic assumption is that protein abundance is proportional to the number of spectra (after normalization)
→ Quantification can also be based on the comparison of features, which can be defined as all mass-spectrometric signals (peaks) caused by the same peptide

Hypothesis-driven, targeted bottom-up proteomics approaches

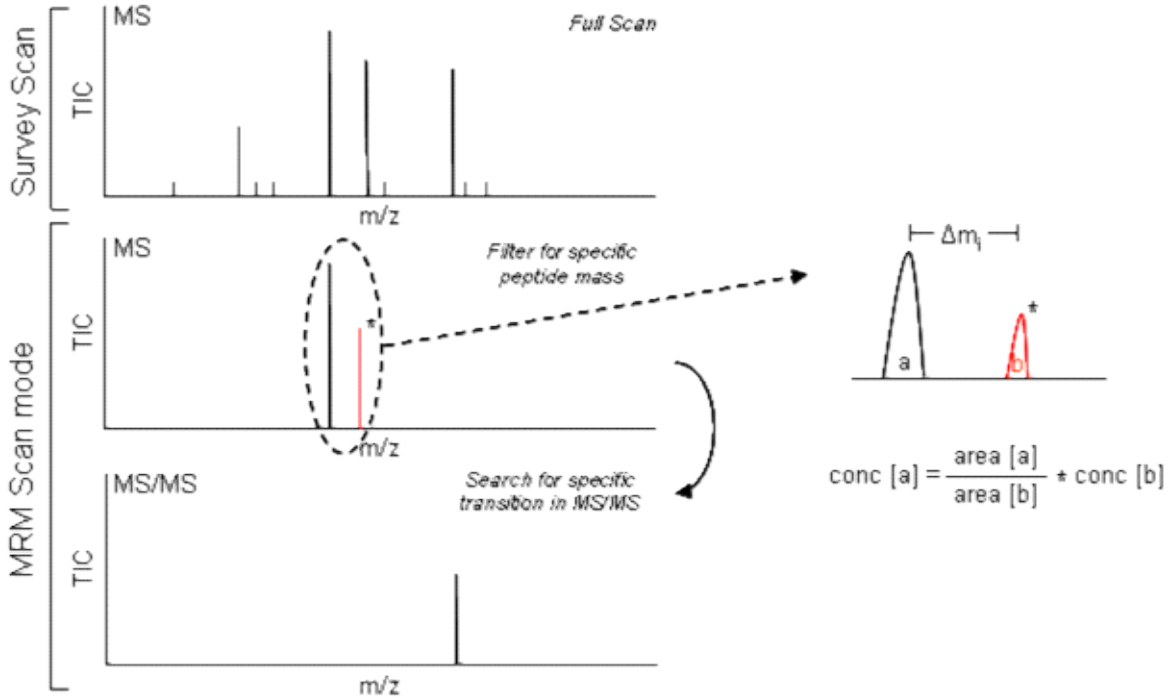


Uzozie & Aebersold, Journal of Proteomics, 2018

S/MRM: Selected/Multiple Reaction Monitoring; the proteins are pre-selected and provide information on the characteristic peptide precursor and fragment ion signals (transitions)

PRM: Parallel Reaction Monitoring; similar to S/MRM, but all resulting fragment ion signals from a precursor ion are monitored

DIA/SWATH: Data Independent Acquisition/Sequential Windowed Acquisition of All Theoretical Mass Spectra



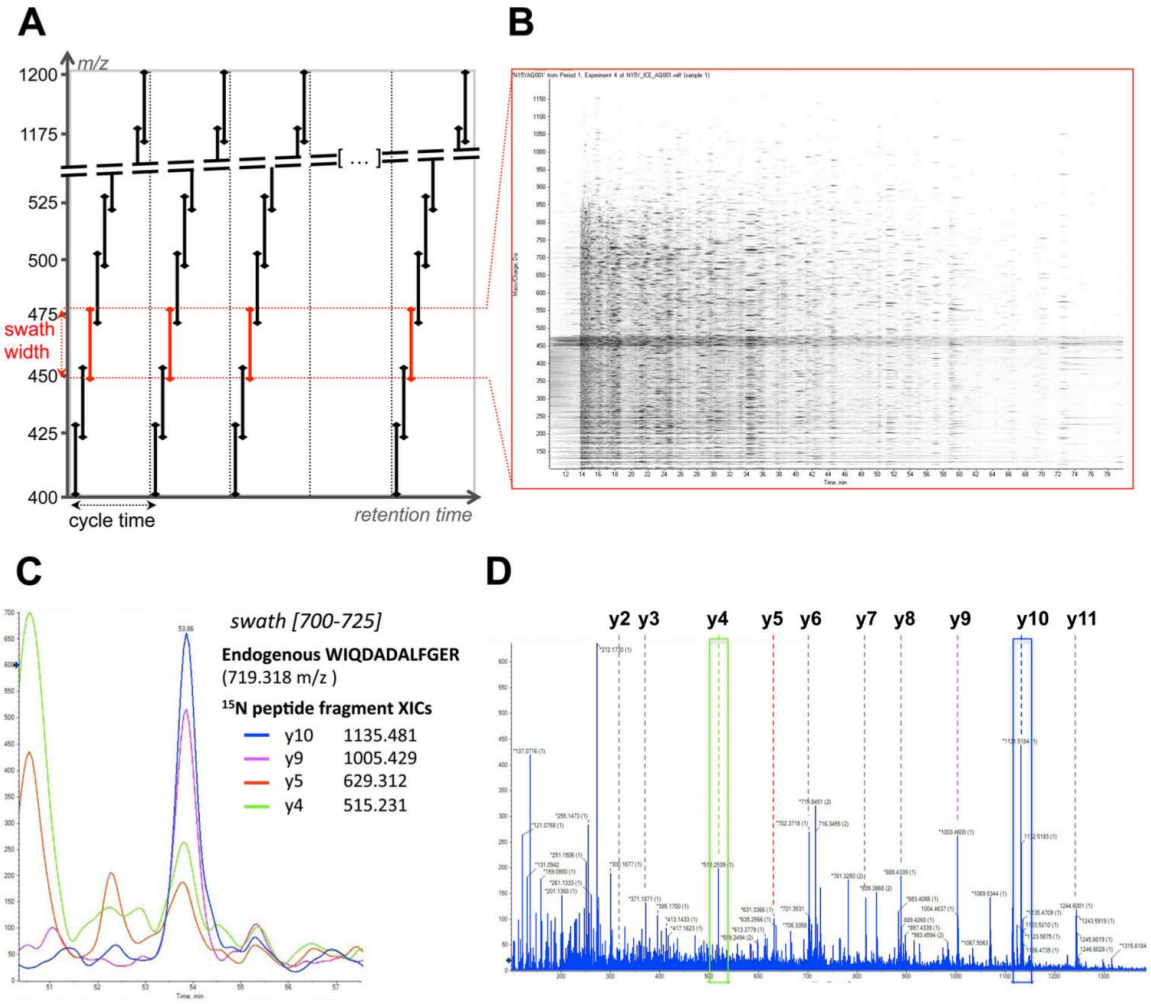
Baginsky, *Mass Spectrometry Reviews*, 2008

=> Requires well detectable peptides that are unique for a given protein

=> Experimental spectra help selecting good transitions

→ In Parallel Reaction Monitoring (PRM), all fragment ion signals from a precursor ion are monitored, achieving better selectivity and better quantitative accuracy

SWATH-MS (data independent analysis DIA)



A) the mass spectrometer steps through a set of precursor acquisition windows

B) In each cycle it fragments all precursors from all the respective quadrupole isolation windows and in each isolation window it records a complete, high accuracy fragment ion spectrum of all precursors

C) The data are analyzed by reconstructing the lineage of precursor and fragment ions based on their chromatographic elution profile, or with software for automated targeted data analysis

@ http://www.imsb.ethz.ch/research/aebersold/research/swath-ms.html