

# Mass spectrometry

- Mass spectrometry
- Database searches
- Result significance
- De novo sequencing

Analyzing MS/MS data to identify  
the corresponding peptide or protein

Matching MS/MS data to protein,  
EST or genomic databases.

# Databases

- Biological databases with sequence information are usually plain ASCII files.
- FASTA is the most common format used.
- Each entry is preceded by a header containing the description.
- Then the actual entry follows, usually 80-100 capitalized letters, terminated with a linefeed character (`\n`).

# Database examples

## Genomic

```
>scaffold_1152
GGTGCGGCCGTCCTCCAGCTGCTTGCCGGCGAAGATCAGGCGCTGCTGGT
CCGGGGGGATGCCTGCATCCGGTGAGGAAACGCTCGTGTGACACAAAGTG
GGTGGGCGCAGGAAGCAGCAATCAACACAGCCCAGTGCAGCTGCAAAGCG
CCCGCCTTACCACTGACCCGCCTGGCCACCCACCCCTACCCCCCGTAAGG
AAAGAGCCCCGACTCACCTCCTTGTCTGAATCTTGGCCTTCACGTTCT
CAATGGTGTCCGAAGACTCCACCTCGAGCGTGATGGTCTTGCCCGTCAGG
GTCTTGACGAAGATCTGCATGCCACCGCGCAGGCGCAGCACCAGGTGCAG
```

## Translated

```
>RF1_scaffold_1152
GAAVLQLLAGEDQALLVRGDACIR$GNARVRQSGWAQEAAINTAQCS
KAPALPLTRLATHPYPP$GKSPDPSLS$ILARDVAHDFAKSSPR$YA
PLIPQNLRC$SIEMKQPASLLSPIGEGACASHLQCLEKCLLP$GAI
VY
MIS$GSGRR$TSWVGIGGCNDGTEKRSEVDSRRGGKGNID
>RF2_scaffold_1152
VRPSSSCLPAKIRRCWSGGMPASGEETLVS AATAAKPQTWSPTA
WEF
KVGGRRKQOSTQPSAAAKRPPYH$PAWPPTPTPRKERAPTHPPC
PESW
SRSQWCPKTPPRA$WSCPSGS$RRSACHRAGAAPGAGSTPSGCC
SQPG
CGRPPAACRRRSGAAGPGGCLCVGGGGEGACASHLQCLEGE
```

# Search methods

- Plain text search
- Mass fingerprinting
- Sequence tags
- Cross correlation
- Statistical approach (usually Bayesian)
- Any combination of the above

# Search method restrictions

- Depending on the preparation
  - protein abundance
  - pureness of the sample (number of proteins)
- Depending on the mass spectrometer
  - ability to fragment peptides
  - mass resolution
- Depending on the identification
  - Search method used
  - Its “synergetic” limitations with the above

# Plain text search

- Comparison of the input string with all strings in the database.
- Possibility of adding various degrees of freedom
  - Conservative substitutions, etc.
- Matching only subsets of the sequence

# Plain text search

- Search improvements
  - Substitution matrices
  - Searching with subsets of the input string
- Problem
  - long processing time (the complete text file must be evaluated with the string in question)
  - Thus processor time demanding

Mackey et al 2002, Amer. Soc. Biochem. MolBiol., 139-147

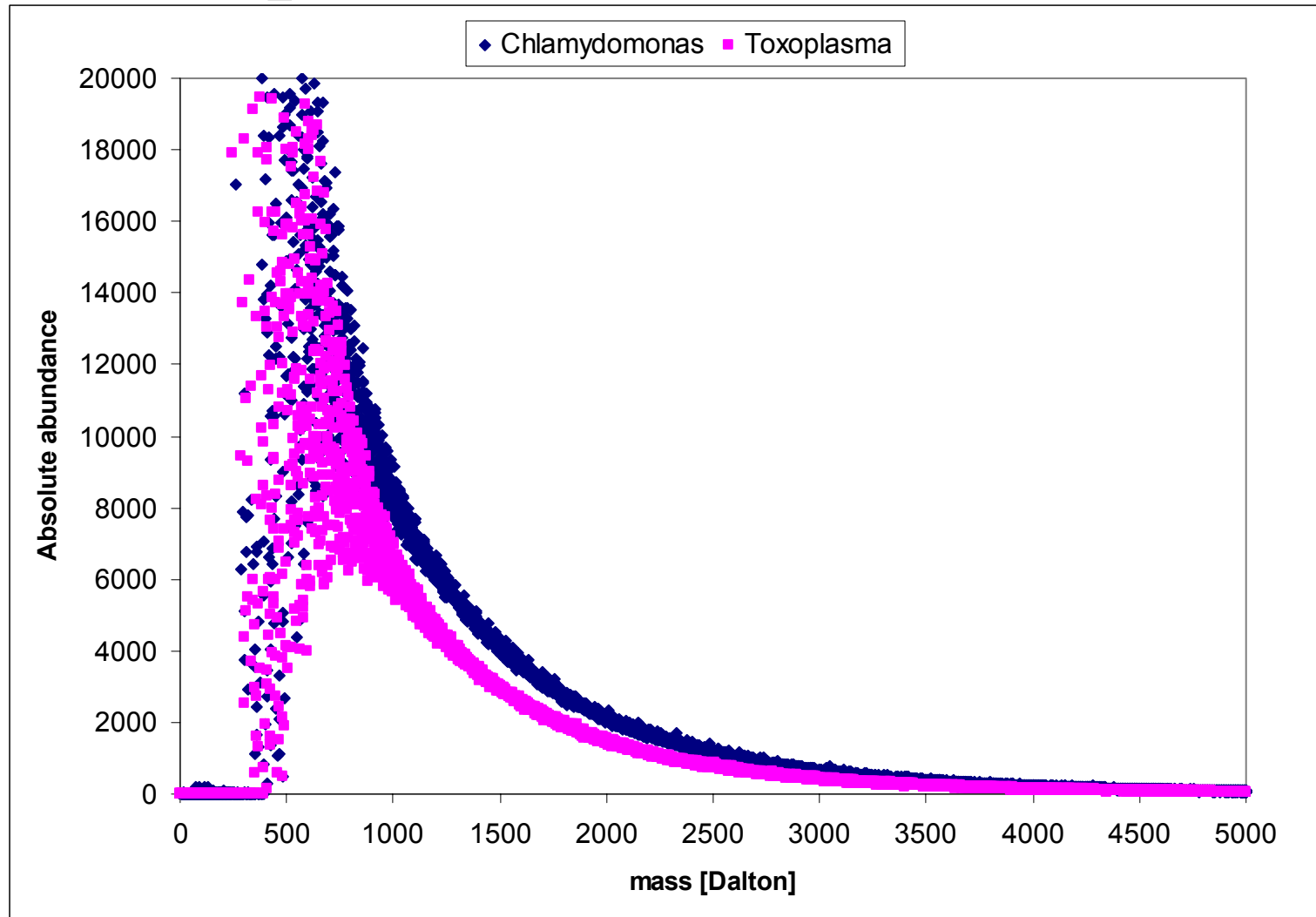
Altschul et al 1997, Nuc. Acid Res., 25, 17, 3389-3402

Pearson and Lipman 1988, Proc. Natl. Acad. Sci., 85, 2444-2448

# Mass fingerprinting

- Digesting the complete database with the enzyme in question.
- Calculating the peptide masses.
- Finding the location with most fitting masses.
- Problems
  - Some masses are quite abundant
  - Not annotated genomes could pose problems

# Peptide mass abundance



The genomic databases were translated in all six reading frames on the fly and then digested with Trypsin (R|K).

# Annotation

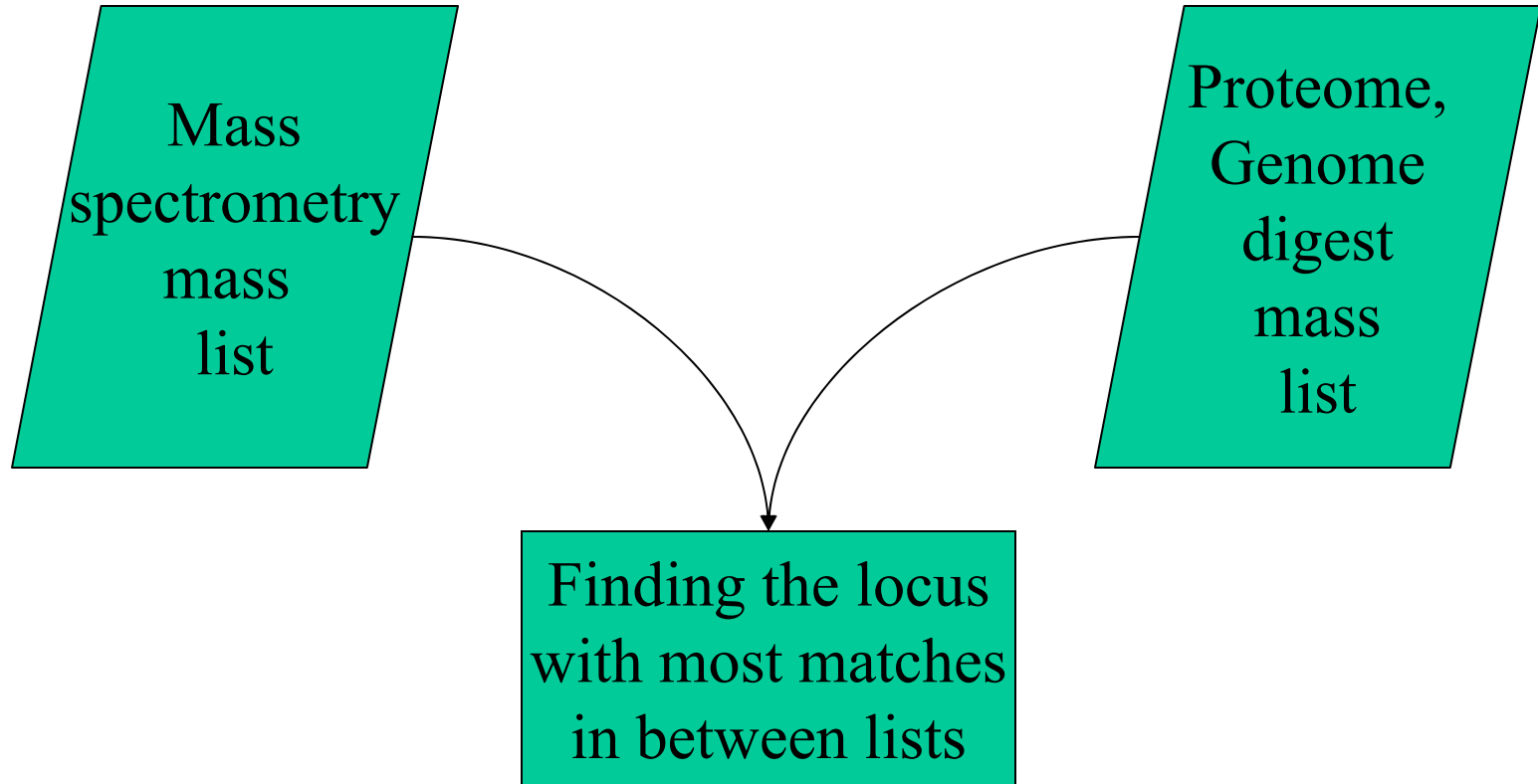
Masses are matched against the database.

If it is correctly annotated, the protein may be identified by finding the protein with the most matching masses.

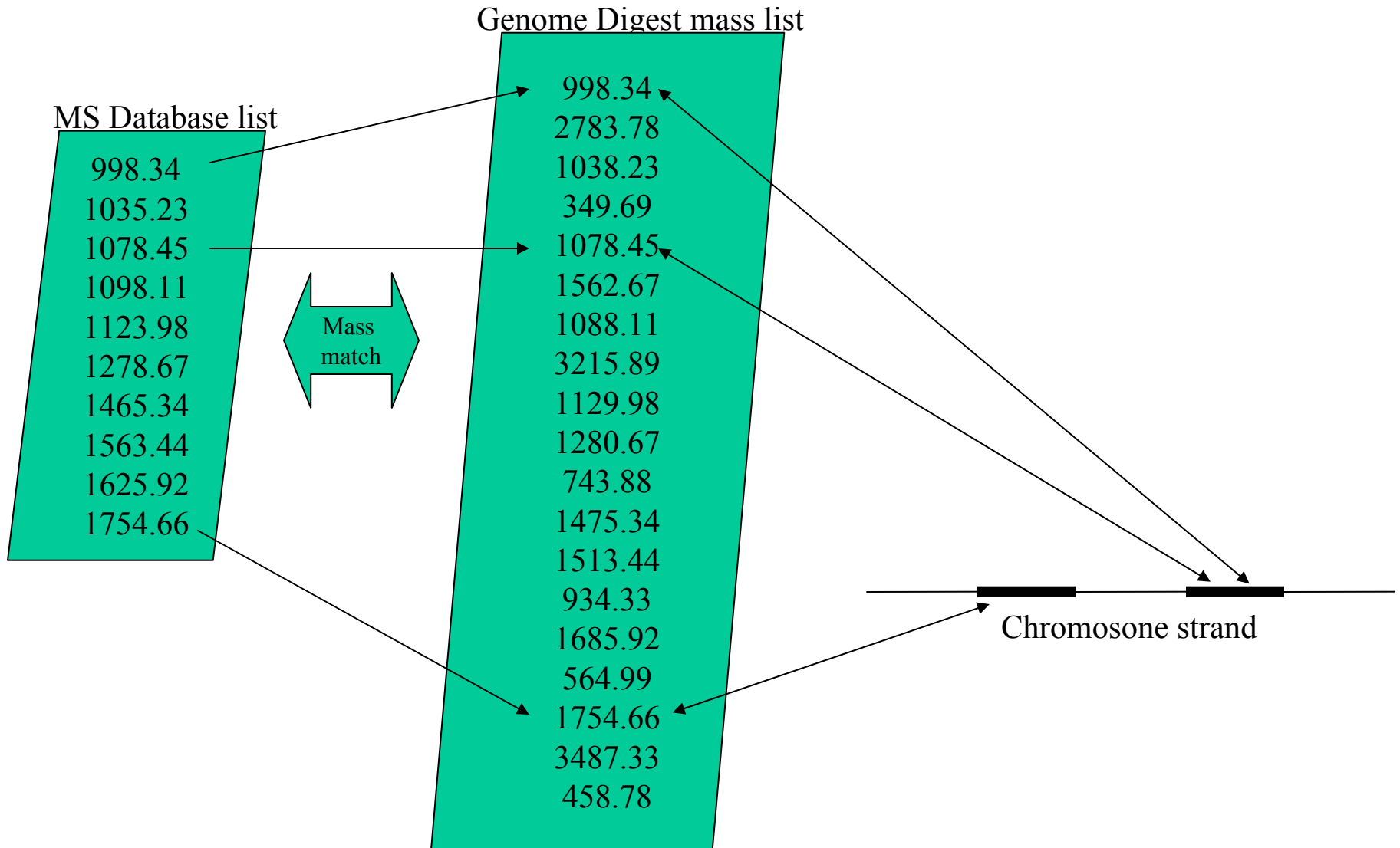
If the database is not annotated or annotation is questionable then defining a window for the search is quite problematic.

If the input is derived from a protein mixture, it is even more complicated.

# Mass fingerprinting



# Mass fingerprinting



# Mass fingerprinting

- Significance
  - absolute mass
  - peptide mixture
  - number of matching peptides

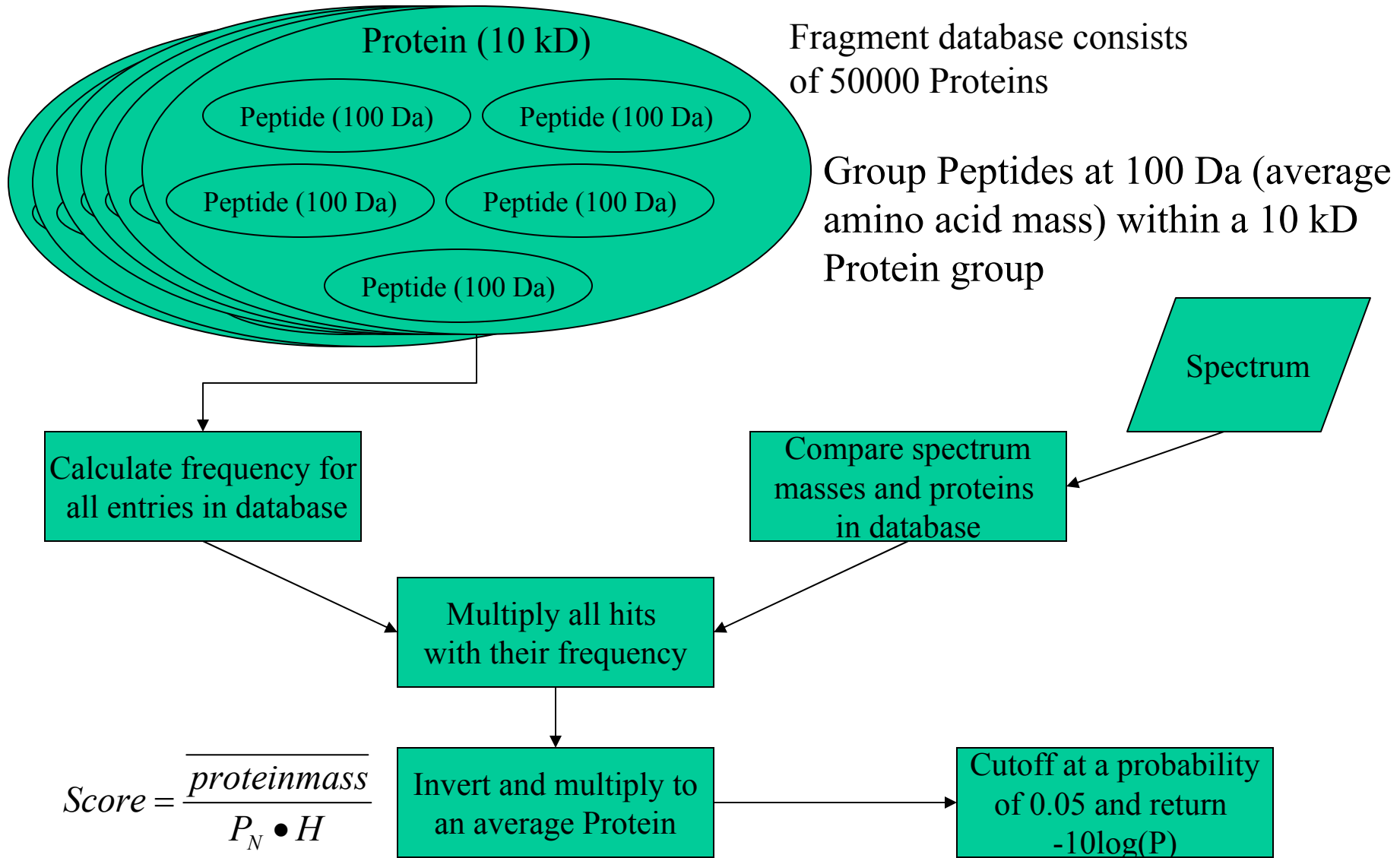
Giddings et al 2002, PNAS, 100, 20-25

Perkins et al 1999, Electrophoresis, 20, 3551-3567

# MoWSe score

- Scoring based on peptide mass frequency
- Algorithm
  - Group Proteins in 10 kDa bins (database)
  - Place Peptides in 100 Da bins (database)
  - Frequency = Peptide bin / total Peptides / Protein
  - Normalize to largest bin value
  - Multiply frequency with each peptide in spectrum
  - Invert, multiply and normalize to a 50000 kDa protein:
    - Score =  $50000 / P_N * H$ 
      - H = MW of matched Protein
      - $P_N$  = Product of frequency scores

# MoWSe score

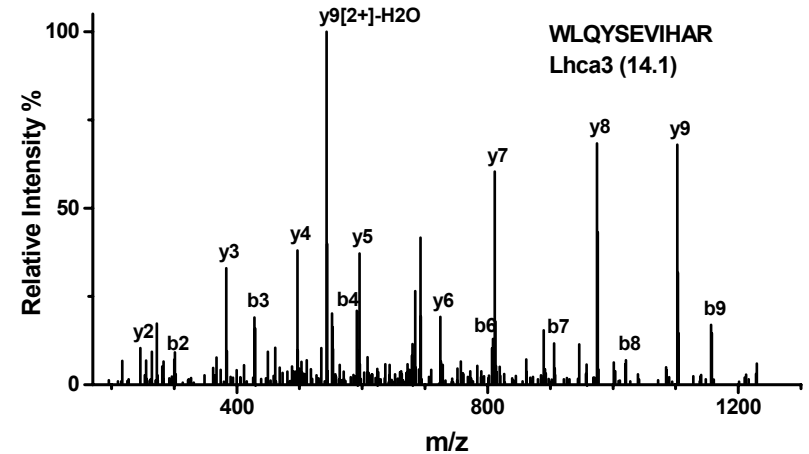


# Sequence Tags

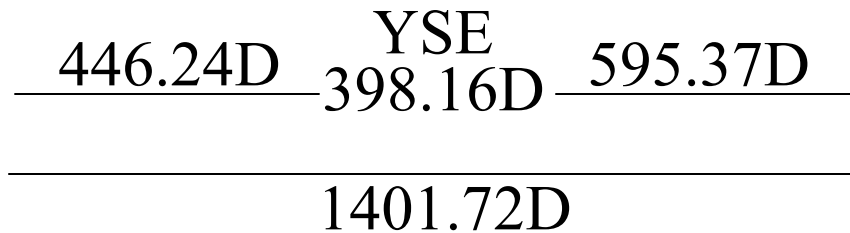
- In addition to mass fingerprinting, short sequences are used here.
- Higher confidence in search results than mere mass fingerprinting
  - Short sequence
  - Position of sequence in peptide
  - Mass and two sub masses

# Sequence Tags

Peptide example: WLQYSEVIHAR



Sequence Tag example: [446.24]YSE[595.37]



# Sequence Tags

- Significance for several matching sequence tags as proposed by Sunyaev et al.
- Several probabilities have to be taken into account
  - N-terminal mass tag matching  $P = \left(1 - e^{-kP_1(m_{1N})f(a_{11})f(a_{12})f(a_{13})(1-P_1(m_{1C}))}\right)$
  - C-terminal mass tag matching  $- \left(1 - e^{-k(1-P_2(m_{2N}))f(a_{21})f(a_{22})f(a_{23})P_2(m_{2C})}\right)$
  - Sequence matching  $- \left(1 - e^{-kP_3(m_{3N})f(a_{31})(1-f(a_{32}))f(a_{33})P_3(m_{3C})}\right)$
  - Precursor mass matching
- E value is calculated by considering all arbitrary peptides that would be generate a better score than the match
- Subtracting result from 1 and multiplying by database size

Tabb et al 2003, Anal. Chem., 75, 6415-6421

Sunyaev et al 2003, Anal. Chem., 75, 1307-1315

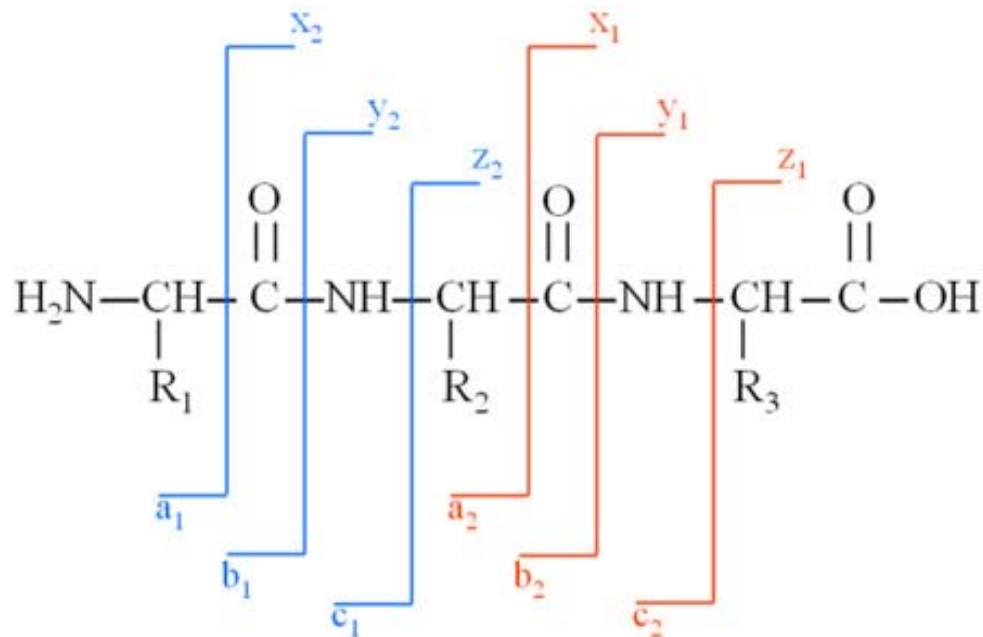
Perkins et al 1999, Electrophoresis, 20, 3551-3567

Mann and Wilm 1994, Anal. Chem., 66 4390-4399

# Cross correlation

- Digesting the database with the enzyme in question.
- Picking all fragments within a mass window close to the precursor mass of the peptide in the mass spectrum
- Calculating an artificial spectrum from all those fragments
- Cross correlate spectra to original mass spectrum

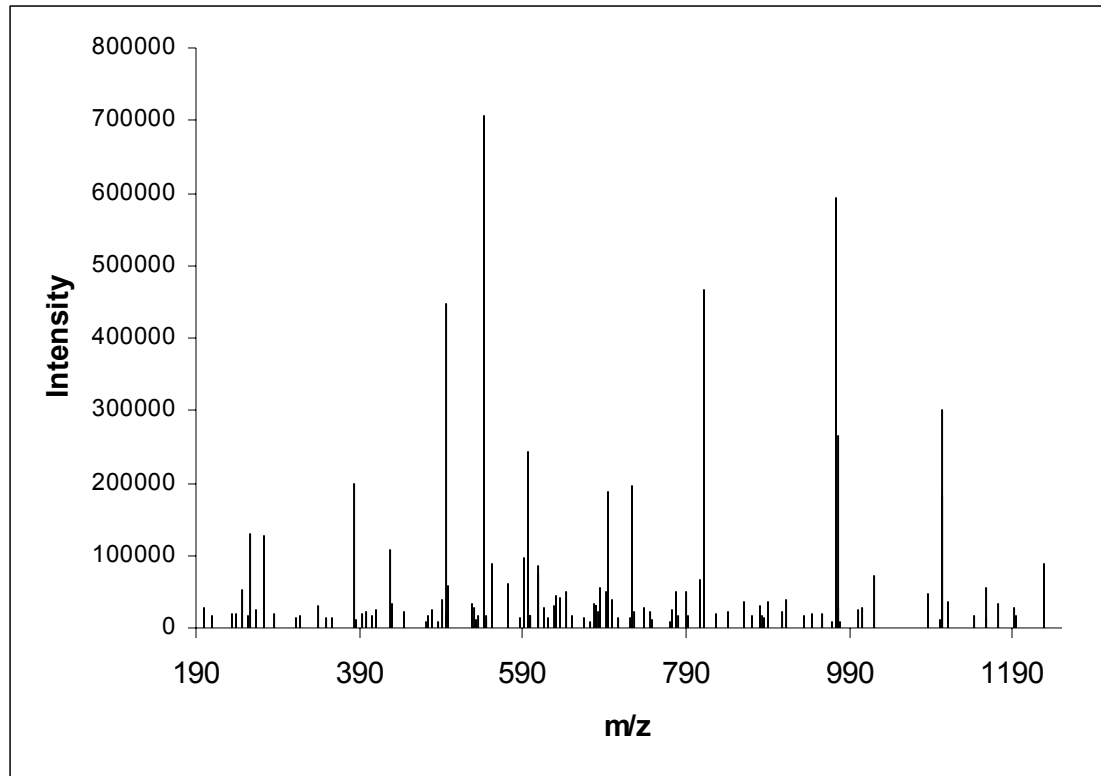
# Collision induced dissociation



Every bond in a peptide can and may break in any CID experiment. However,  $y$  and  $b$ -ions form the most prominent ion fraction in practice.

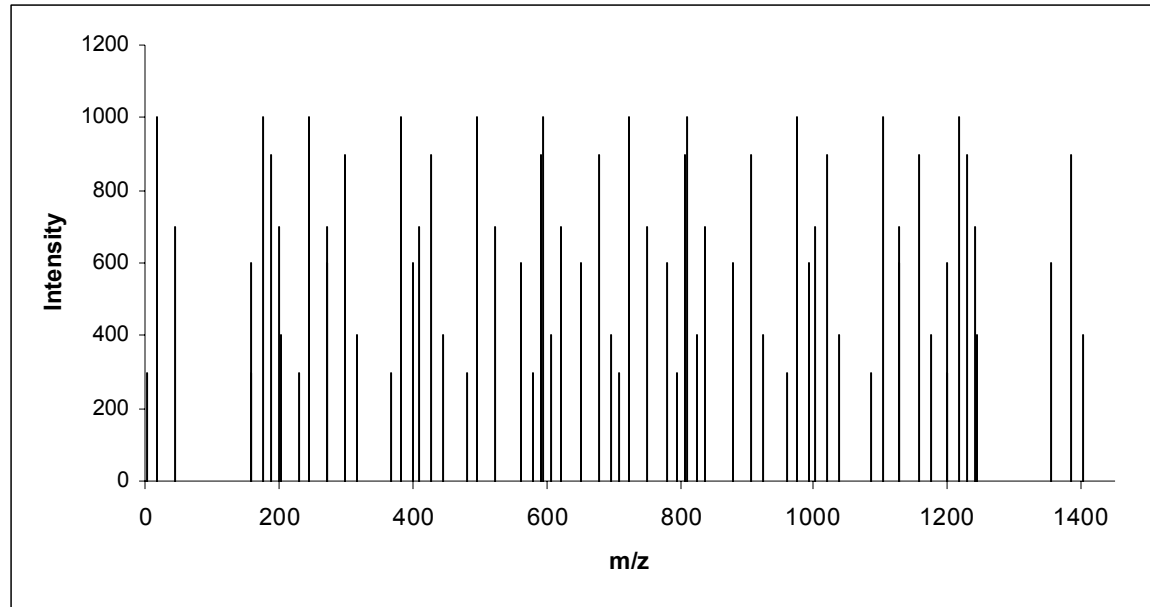
This is dependent on the collision energy used. The higher this energy the more bonds break.

# An example mass spectrum



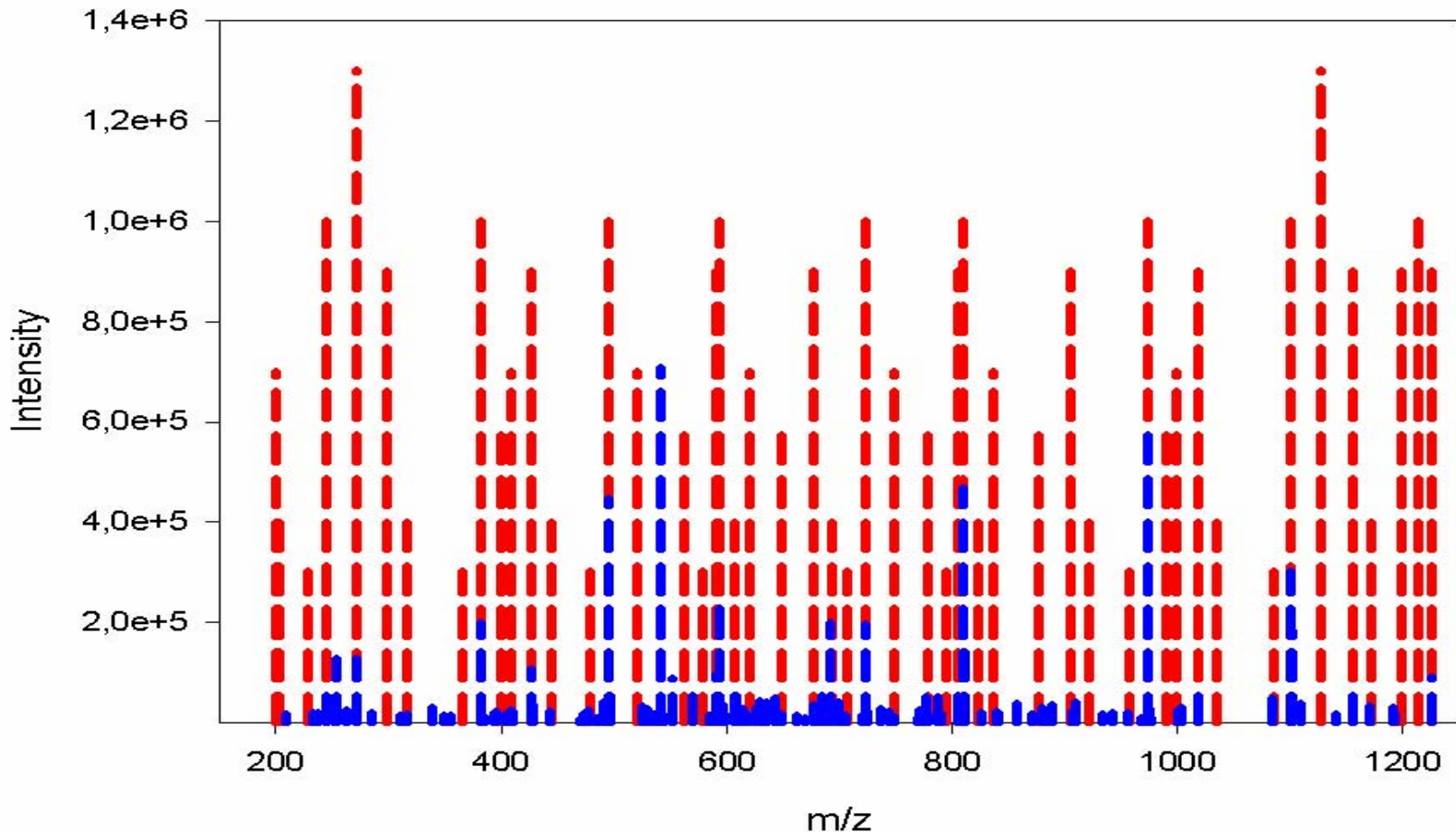
The spectrum of the peptide WLQYSEVIHAR after fragmentation by collision induced dissociation.

# An example for a theoretical spectrum



Theoretical mass spectrum of the peptide with the amino acid sequence WLQYSEVIHAR. y-ions set to an intensity of 1000i, b-ions 900i, x-ions 700i, a-ions 600i, c-ions 400, and z-ions 300i, arbitrarily. This was done to be able to graphically differentiate between the ions. Masses found several times were added up in their intensity.

# Calculating an artificial spectrum



Blue spectrum is practically measured and red spectrum is artificially calculated (y,b,x,c,a,z ions). Intensities in the artificial spectrum are arbitrarily chosen.

# Cross correlation

$$r(d) = \frac{\sum_i [(x(i) - \bar{x}) \cdot (y(i - d) - \bar{y})]}{\sqrt{(x(i) - \bar{x})^2} \sqrt{(y(i - d) - \bar{y})^2}}$$

$\underline{x}$ : Spectrum x

$\bar{x}$ : Mean of spectrum x

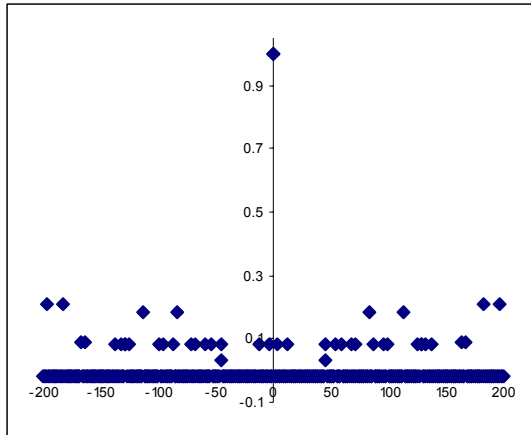
$\bar{y}$ : Mean of spectrum y

y: Spectrum y

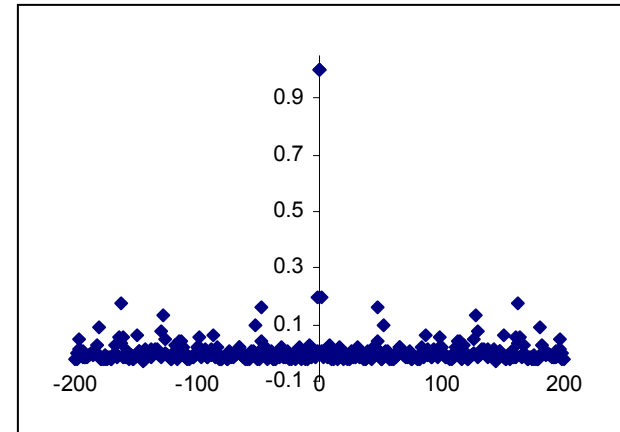
d: delay of one spectrum to the other

i: index

# Perfect cross correlation

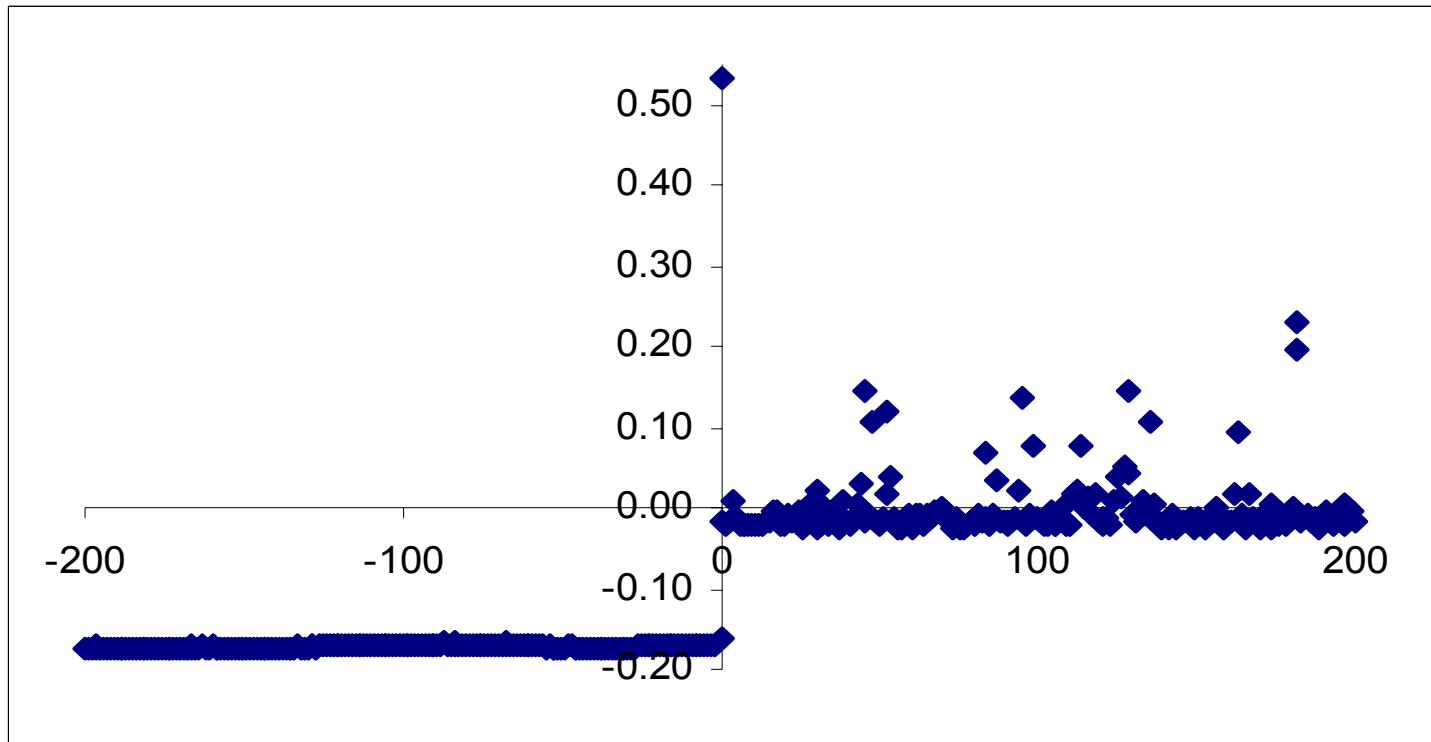


Cross correlation of the theoretical spectrum with itself (auto correlation)



Cross correlation of the practical spectrum with itself (auto correlation)

# Comparative cross correlation



Cross correlation of the theoretical and the measured spectrum for a delay of +/- 200. Not normalized to the auto correlation. Only b and y-ions used in calculation.

# Fast Fourier transformation

- Various FFT's have been implemented
  - Radix-2, Radix-4, Split-Radix
  - Fast Hartley Transform (FHT)
  - Quick Fourier Transform (QFT)
  - Decimation in time frequency (DIFT)
- Reduces complexity from  $O(N^2)$  to  $O(N \log N)$
- FHT is the most efficient algorithm (Ganapathiraju)

Analysis and Characterization of Fast Fourier Transform Algorithms  
Aravind Ganapathiraju  
Institute for Signal and Information Processing  
Mississippi State University

# Fourier Transform

A technique for expressing a waveform as a weighted sum of sines and cosines. Fast Fourier transform further reduces the number of computations necessary by placing restrictions on N  
Fourier and Laplace transformations are widely used in science.

## Fast Hartley Transform

The maybe most efficient implementation of an FFT algorithm

$$X(k) = \frac{1}{\sqrt{N}} \cdot \sum_{n=0}^{N-1} x(n) \cdot \left[ \cos\left(\frac{2\pi kn}{N}\right) + \sin\left(\frac{2\pi kn}{N}\right) \right]$$

# Fast Fourier transformation

- Reduces the number of necessary calculations
- Speeds up the overall search
- Questionable significance threshold
  - Significant XCorr (Sequest)
    - >1.5 for singly charged ions
    - >2.5 for doubly charged ions
    - >3.5 for triply charged ions

# Statistical approach

- Bayesian statistics
  - Known factors (common sense) can be easily incorporated in the scoring function
  - inference from incomplete information
- New, promising research field

Zhang et al 2002, Proteomics, 2, 1406-1412

Zhang and Chait 2000, Anal. Chem., 72, 2482-2489

# Database search programs

	Text search	preucursor mass(es)	fragment masses	Cross correlation	Stochastic models	Spectral Product	Dot Product
<b>Sequest</b>		+	+	?			?
<b>Mascot</b>	+	+	+				
<b>GPF</b>	+	+					
<b>Pepsea</b>							
<b>Pepsearch</b>							
<b>MS-Blast</b>	+	+	+				
<b>FASTA</b>	+						
<b>PeptIdent</b>		+					
<b>Sherpa</b>		+					
<b>Blast</b>	+						
<b>GutenTag</b>		+	+				
<b>MultiTag</b>		+	+				
<b>ProBID</b>					+		
<b>ProFound</b>					+		
<b>OpenSea</b>							
<b>ProteinLynx</b>							
<b>MS-Shotgun</b>							
<b>CIDentify</b>							
<b>AutoMod</b>							
<b>pFind</b>							+
<b>Sonar</b>							+
<b>SCOPE</b>					+		
<b>Sherenga</b>					+		

# Database search Programs

Allmer et al, 2004, FEBS Letters, 28287, 1-5

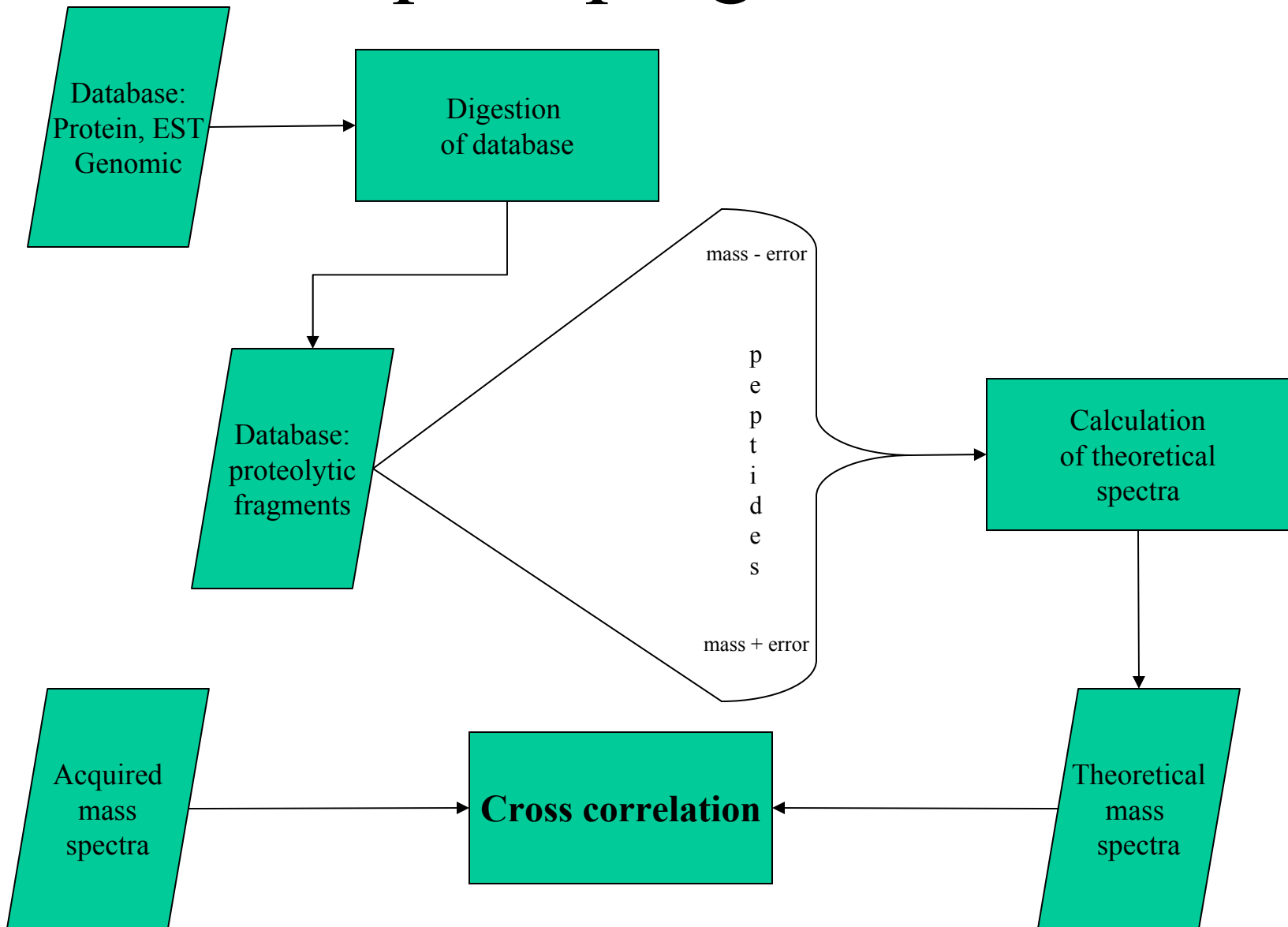
Eng et al 1994, Amer. Soc. Mass Spec., 94, 1044-0305

Perkins et al 1999, Electrophoresis, 20, 3551-3567

# Sequest general information

- Thermo Electron
- Ion count
- Cross correlation
- EST and translated Genome

# Sequest program flow



# Sequest:

Eng J, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**:976-989

Step 1: Data reduction of the MS/MS spektrum

Step 2: Peptide, in a given database that match with the measured mass are selected.

Step 3: Measured fragment ions and theoretical fragment ion are compared and the best 500 sequences are selected

Step 4: A correlation analysis is conducted to determine scores and rankings

# Cross correlation analysis

- Reconstruction of a theoretical MS/MS spektrum taken from a selected amino acid sequence from the database: predicted m/z- values and magnitude (empirical knowledge, b- and y-type ions x 50; neutral loss of ammonium, water, a-type ions x 10)
- Original spektrum: Deletion of the precursor ion, divide in 10 regions, normalisation to 50

# Advantages:

- Effective use of fragmentation pattern for the identification
- No additional information required (no manual interpretation)

# Disadvantages:

- False-positive or -negative identifications possible
- Methods relies on the presence of the respective peptide sequence in the database

# Database entry

>sp|P02769|ALBU\_BOVIN Serum albumin precursor (Allergen Bos d 6) - Bos taurus (Bovine).  
MKWVTFISLLLLFSSAYSRGVFRRDTHKSEIAHRFKDLGEEHFKGLVLIAFSQYLQQCPF  
DEHVKLVNELTEFAKTCVADESHAGCEKSLHTLFGDELCKVASLRETYGDMADCCEKQEP  
ERNECFLSHKDDSPDLPKLPDPNTLCDEFKADEKFKFWGKYLYEIARRHPYFYAPELLYY  
ANKYNGVFQECCQAEDKGACLLPKIETMREKVLASSARQRLRCASIQKFGERALKAWSVA  
RLSQKFPKAEFVEVTKLVTDLTKVHKECCHGDLLECADDRADLAKYICDNQDTISSKLKE  
CCDKPLLEKSHCIAEVEKDAIPENLPPLTADFAEDKDVCKNYQEAKDAFLGSFLYEYSRR  
HPEYAVSVLLRLAKEYEATLEECCA KDDPHACYSTVFDK LKHLVDEPQNLIKQNC DQFEK  
LGEYGFQNALIVRYTRKVPQVSTPTLVEVSRS LGKVGTRCCTKPESERMPCTEDYLSLIL  
NRLCVLHEKTPVSEKVTKCCTESLVNRRPCFSALTPDETYVPKAFDEKLFTFHADICTLP  
DTEKQIKKQTALVELLKHKPKATEEQ LKTVMENFVAFVDKCCAADDKEACFAVEGP KLVV  
STQTALA

# Sequest results

**Sequest Summary - Netscape**

Datei Bearbeiten Ansicht Gehe Communicator Hilfe

## Sequest Summary

[Setup](#) [CreateDTA](#) [FuDTA](#) [RunSequest](#) [Status](#) [Summary](#) [Utilities](#) [Home](#)

Sample: Schubert, R. (BSA\_800) BSA rs      Db: [bsa \(07/16/2002\)](#)      Inspector [View Info](#)      Mass: avg  
Datafiles: [RS\\_Test01 \(11/13/2002-11/13/2002\)](#)      Dir: [rschubertbsa 800](#)      Enz: [Trypsin](#)      Tot: 285|260|1.3e9  
Intensity:  Full  Zoom  MS2      Diff Mods: 0.000 C 0.000 S X

Consensus Chron Max rank:  Max list:  Controls: [Show](#) [DTA VCR](#)

#	TIC	File	z	dM	MH+	%corr	dCn	Sp	RSp	Ions	Ref	( )	Sequence	Pull to Top
110	1.5e7	<a href="#">1293-1296</a>	3	1.0	1439.7	4.77	0.00	3214	1	31/44	<a href="#">sp p02769 albu</a>	(R)	<a href="#">RHPEYAVSVLLR</a>	
225	1.8e7	<a href="#">2103-2106</a>	2	1.0	1567.7	4.73	0.00	2127	1	19/24	<a href="#">sp p02769 albu</a>	(K)	<a href="#">DAFLGSFLYEYSR</a>	
214	1.6e7	<a href="#">1914-1917</a>	2	0.9	1399.7	4.58	0.00	1545	1	19/22	<a href="#">sp p02769 albu</a>	(K)	<a href="#">TMENFVAFVDK</a>	
185	1.7e7	<a href="#">1719-1722</a>	2	0.9	1479.8	4.56	0.00	1332	1	17/24	<a href="#">sp p02769 albu</a>	(K)	<a href="#">LGEYGFQNALIVR</a>	
56	4.9e7	<a href="#">0663-0666</a>	3	0.7	1249.7	4.07	0.00	2330	1	26/36	<a href="#">sp p02769 albu</a>	(R)	<a href="#">FKDLGEEHFK</a>	
112	2.0e7	<a href="#">1311-1314</a>	2	0.6	1142.8	3.69	0.00	1091	1	15/18	<a href="#">sp p02769 albu</a>	(K)	<a href="#">KQTALVELLK</a>	
57	1.0e7	<a href="#">0669-0672</a>	2	0.8	1249.6	3.61	0.00	1236	1	16/18	<a href="#">sp p02769 albu</a>	(R)	<a href="#">FKDLGEEHFK</a>	
120	1.8e7	<a href="#">1356-1365</a>	2	1.0	1639.9	3.46	0.00	656	1	18/28	<a href="#">sp p02769 albu</a>	(R)	<a href="#">KVPQVSTPTLVEVSR</a>	
98	7.9e5	<a href="#">1215</a>	3	0.9	1546.9	3.17	0.00	1288	1	22/48	<a href="#">sp p02769 albu</a>	(K)	<a href="#">LKHLVDEPQNLIK</a>	
147	1.2e7	<a href="#">1512-1515</a>	2	0.9	1511.8	3.13	0.00	1220	1	20/26	<a href="#">sp p02769 albu</a>	(K)	<a href="#">VPOVSTPTLVEVSR</a>	
153	1.1e7	<a href="#">1530-1533</a>	2	0.9	1163.4	3.04	0.00	1018	1	16/18	<a href="#">sp p02769 albu</a>	(K)	<a href="#">LVNELTEFAK</a>	
102	6.9e6	<a href="#">1251-1254</a>	2	0.8	1305.8	3.00	0.00	1050	1	16/20	<a href="#">sp p02769 albu</a>	(K)	<a href="#">HLVDEPQNLIK</a>	
143	1.2e7	<a href="#">1494-1500</a>	2	0.7	1002.5	2.99	1.00	596	1	14/18	<a href="#">sp p02769 albu</a>	(K)	<a href="#">LVVSTQTALA</a>	
74	1.1e7	<a href="#">0813-0816</a>	2	0.7	922.4	2.82	0.00	613	1	13/14	<a href="#">sp p02769 albu</a>	(K)	<a href="#">AEFVEVTK</a>	
50	1.7e7	<a href="#">0630-0642</a>	2	0.6	974.5	2.58	0.00	609	1	13/14	<a href="#">sp p02769 albu</a>	(K)	<a href="#">DLGEEHFK</a>	
15	4.8e7	<a href="#">0393-0450</a>	2	1.3	846.7	2.56	0.00	652	1	10/12	<a href="#">sp p02769 albu</a>	(R)	<a href="#">LSQKFPK</a>	
76	8.9e6	<a href="#">0825-0828</a>	1	0.6	922.4	2.23	0.00	637	1	9/14	<a href="#">sp p02769 albu</a>	(K)	<a href="#">AEFVEVTK</a>	
114	3.8e7	<a href="#">1326-1329</a>	2	0.3	927.7	2.21	0.00	521	1	10/12	<a href="#">sp p02769 albu</a>	(K)	<a href="#">YLVEIAR</a>	
23	4.6e7	<a href="#">0444-0459</a>	2	1.0	817.9	2.12	1.00	502	1	11/12	<a href="#">sp p02769 albu</a>	(K)	<a href="#">ATEEQLK</a>	
163	3.3e7	<a href="#">1584-1587</a>	2	0.8	1014.4	2.10	0.00	219	1	10/16	<a href="#">sp p02769 albu</a>	(K)	<a href="#">QTALVELLK</a>	
51	2.2e7	<a href="#">0633-0636</a>	2	0.7	886.2	1.82	0.00	498	1	10/14	<a href="#">sp p02769 albu</a>	(K)	<a href="#">DDSPDLPK</a>	
27	2.8e6	<a href="#">0468</a>	1	0.5	818.4	1.79	1.00	395	1	9/12	<a href="#">sp p02769 albu</a>	(K)	<a href="#">ATEEQLK</a>	
156	1.4e7	<a href="#">1548-1554</a>	1	0.7	1163.6	1.77	0.00	335	1	10/18	<a href="#">sp p02769 albu</a>	(K)	<a href="#">LVNELTEFAK</a>	
25	3.9e6	<a href="#">0462</a>	1	0.5	609.2	1.68	0.00	433	1	6/8	<a href="#">sp p02769 albu</a>	(K)	<a href="#">AFDEK</a>	
129	2.9e6	<a href="#">1410-1464</a>	2	1.0	1283.5	1.66	0.00	121	1	9/20	<a href="#">sp p02769 albu</a>	(R)	<a href="#">HPEYAVSVLLR</a>	
52	1.4e7	<a href="#">0639-0645</a>	1	0.5	886.4	1.61	0.00	385	1	10/14	<a href="#">sp p02769 albu</a>	(K)	<a href="#">DDSPDLPK</a>	
115	1.9e7	<a href="#">1332-1335</a>	1	0.6	927.5	1.58	0.00	226	1	8/12	<a href="#">sp p02769 albu</a>	(K)	<a href="#">YLVEIAR</a>	
16	1.5e6	<a href="#">0399-0414</a>	1	0.0	752.8	1.54	0.00	173	1	7/10	<a href="#">sp p02769 albu</a>	(K)	<a href="#">NYQEAK</a>	
62	1.9e7	<a href="#">0708-0711</a>	1	0.5	789.5	1.54	0.00	437	1	8/12	<a href="#">sp p02769 albu</a>	(K)	<a href="#">LVTDLTK</a>	
14	2.2e7	<a href="#">0375-0378</a>	1	0.4	660.3	1.53	0.00	211	1	7/10	<a href="#">sp p02769 albu</a>	(K)	<a href="#">TPVSEK</a>	
172	5.9e6	<a href="#">1644-1647</a>	3	-0.6	2035.9	1.37	0.00	196	1	19/76	<a href="#">sp p02769 albu</a>	(K)	<a href="#">EACFAVEGPKLVVSTQTALA</a>	
12	4.8e6	<a href="#">0363-0381</a>	1	-0.8	518.4	1.34	0.00	308	1	6/8	<a href="#">sp p02769 albu</a>	(R)	<a href="#">ADLAK</a>	

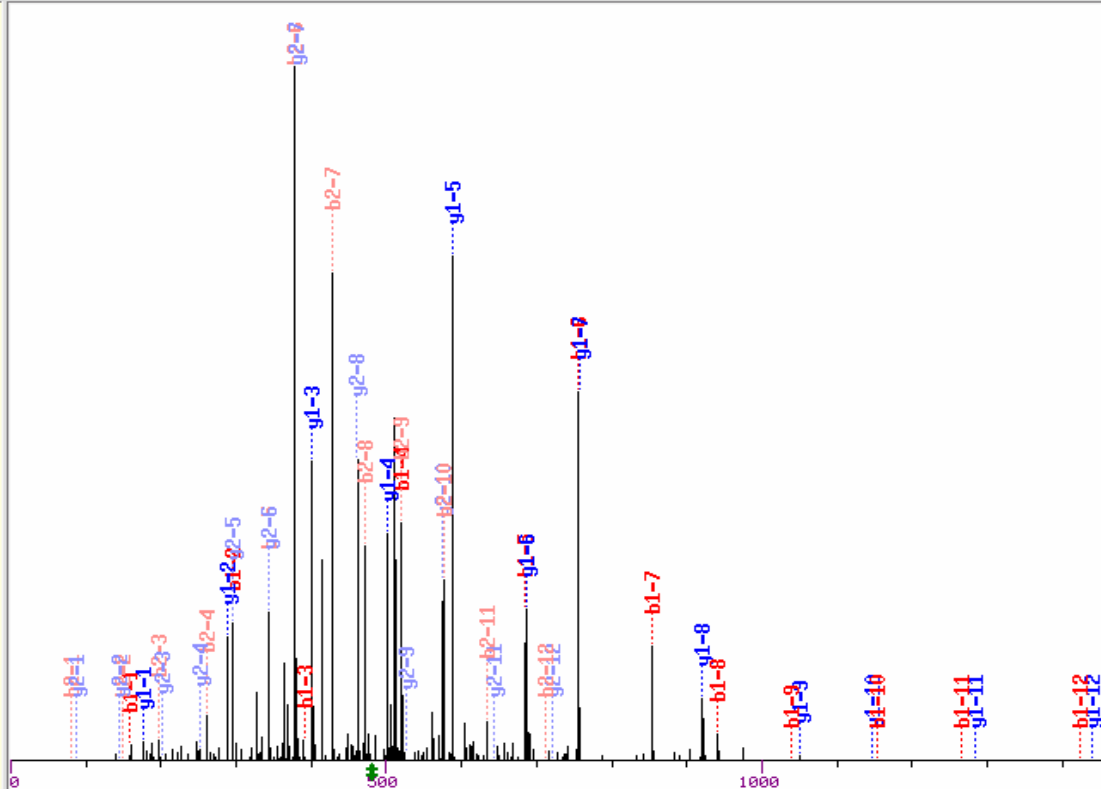
Dokument: Übermittelt

© UW Biological Mass Spectrometry Lab

[1-axis](#) [2-axis](#) [3-axis](#) [4-axis](#) [5-axis](#) [Control](#) [Full View](#) [Ion List](#) [Zoom](#) [1296](#) [Fuzzy](#)

Seq #	b	y	(+1)
R 1	157.2	1440.7	12
H 2	294.3	1284.5	11
P 3	391.5	1147.4	10
E 4	520.6	1050.2	9
Y 5	683.7	921.1	8
A 6	754.8	758.0	7
V 7	854.0	686.9	6
S 8	941.0	587.7	5
V 9	1040.2	500.7	4
L 10	1153.3	401.5	3
L 11	1266.5	288.4	2
R 12	1422.7	175.2	1

Seq #	b	y	(+2)
R 1	79.1	720.8	12
H 2	147.7	642.8	11
P 3	196.2	574.2	10
E 4	260.8	525.6	9
Y 5	342.4	461.1	8
A 6	377.9	379.5	7
V 7	427.5	343.9	6
S 8	471.0	294.4	5
V 9	520.6	250.8	4
L 10	577.2	201.3	3
L 11	633.7	144.7	2
R 12	711.8	88.1	1



gd 1.2 © 1994, 1995, Quest Protein Database Center, Cold Spring Harbor Labs

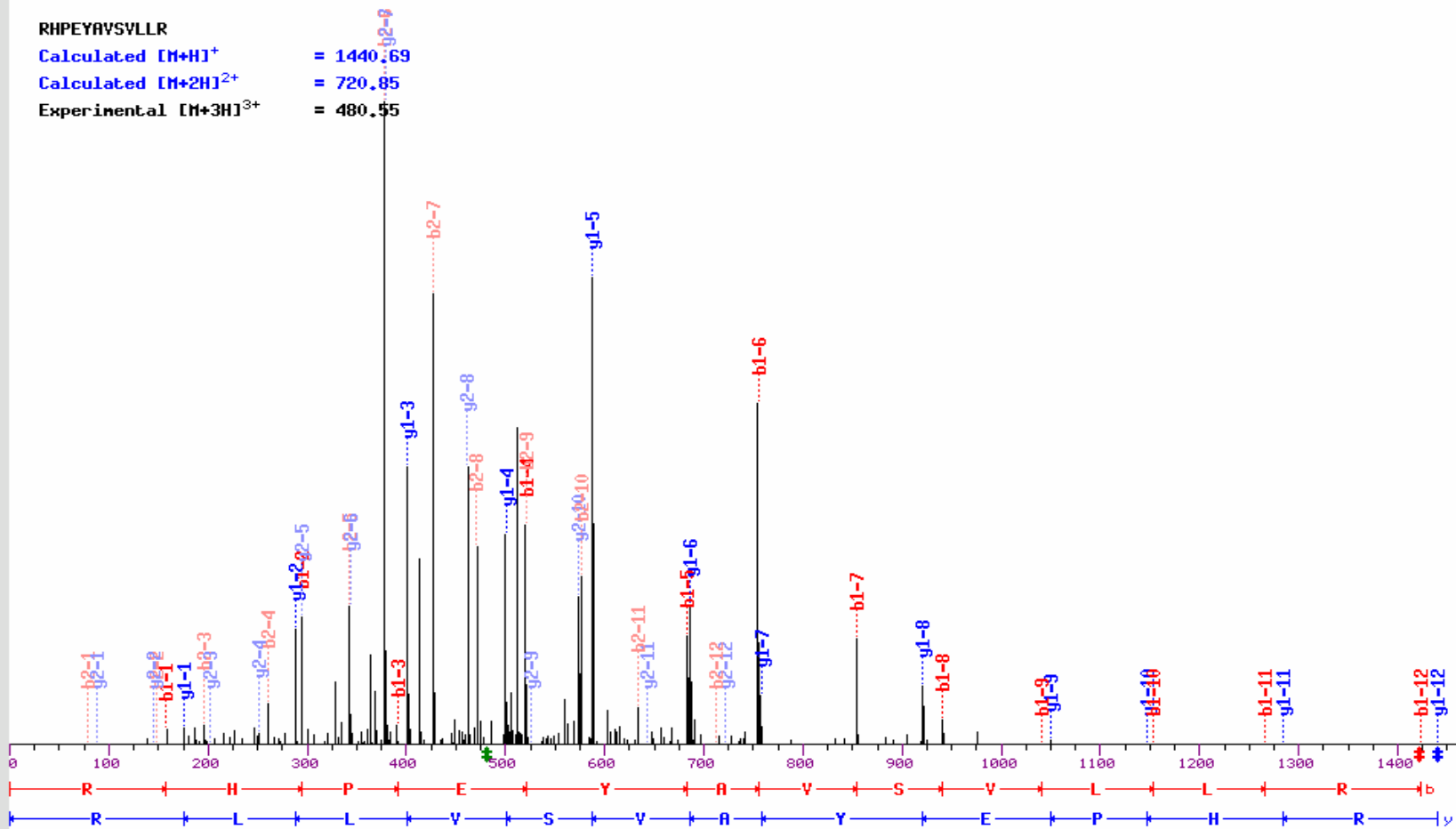
datafile=D:/Sequest/rschubertbsa\_800/RS\_Test01.1293.1296.3.dta, peptide=[RHPEYAVSVLLR](#), Average masses

RHPEYAVSVLLR

Calculated [M+H]<sup>+</sup> = 1440.69

Calculated [M+2H]<sup>2+</sup> = 720.85

Experimental [M+3H]<sup>3+</sup> = 480.55



Ion List [1-axis](#) [2-axis](#) [3-axis](#) [4-axis](#) [5-axis](#)

RS\_Test01.1293.1296.3.out

TurboSEQUENT v.27 (rev. 11), (c) 1999-2000  
 Molecular Biotechnology, Univ. of Washington, J.Eng/J.Yates  
 Licensed to ThermoQuest Corp.  
 11/13/2002, 04:00 PM, 4 sec. on LCQ  
 (M+H)+ mass = 1439.6500 ~ 1.5000 (+3), fragment tol = 0.0, AVG/AVG  
 total inten = 7652.7, lowest Sp = 556.6, # matched peptides = 110157  
 # amino acids = 1368186, # proteins = 0, d:/database/nr.fasta, d:/database/nr.fasta.bin  
 ion series nABY ABCDVWXYZ: 1 1 1 0.0 1.0 0.0 0.0 0.0 0.0 1.0 0.0  
 display top 12/12, ion % = 0.0, CODE = 10103  
 Enzyme:Trypsin (2)

#	Rank/Sp	(M+H)+	deltCn	XCorr	Sp	Ions	Reference	Peptide
1.	1 / 1	1439.8120	0.0000	4.7720	3213.5	31/44	<a href="#">gi 113582 sp P14639 </a>	+4 (-) RHPEYAVSVLLR
							<a href="#">gi 1351907 sp P02769</a>	
							<a href="#">gi 2190337 emb CAA41</a>	
							<a href="#">gi 229552 prf  75492</a>	
							<a href="#">gi 418694 pir  ABBOS</a>	
2.	2 / 432	1439.7970	0.3880	2.9204	578.1	20/48	<a href="#">gi 18599827 ref XP_0</a>	(-) LLFGGVVFSMIEK
3.	3 / 12	1439.6880	0.3984	2.8707	1163.3	22/48	<a href="#">gi 20890288 ref XP_1</a>	(-) SAQDTPRPSSEHK
4.	4 / 172	1440.7700	0.4196	2.7698	718.3	20/52	<a href="#">gi 22128700 gb AAM92</a>	(-) DVVLGLASPSTEPR
5.	5 / 21	1438.8929	0.4264	2.7374	1078.9	24/48	<a href="#">gi 20890144 ref XP_1</a>	(-) QMNVVPAALKVVR
6.	6 / 61	1438.7729	0.4271	2.7340	870.7	21/52	<a href="#">gi 4140371 gb AAD037</a>	(-) AIYMSGGATSVLLR
7.	7 / 13	1438.7729	0.4277	2.7313	1162.3	24/48	<a href="#">gi 10954995 ref NP_0</a>	+3 (-) SDLGKALAYMLTR
							<a href="#">gi 16264268 ref NP_4</a>	
							<a href="#">gi 16264412 ref NP_4</a>	
							<a href="#">gi 4928615 gb AAD336</a>	
8.	8 / 317	1440.7670	0.4313	2.7138	622.9	18/44	<a href="#">gi 6320037 ref NP_01</a>	(-) ADERRQPVSNLNR
9.	9 / 185	1440.8040	0.4339	2.7013	706.3	18/44	<a href="#">gi 15228832 ref NP_1</a>	(-) QENVKRVNANLR
10.	10 / 8	1440.7450	0.4370	2.6865	1279.6	24/56	<a href="#">gi 21231352 ref NP_6</a>	+1 (-) DDGGLAGPAVKGLDR
							<a href="#">gi 21242675 ref NP_6</a>	
11.	11 / 83	1440.6650	0.4427	2.6593	820.0	21/48	<a href="#">gi 7485718 pir  T051</a>	(-) TKPSSSVVMVMDCR
12.	12 / 31	1439.8220	0.4430	2.6580	994.5	22/52	<a href="#">gi 17564580 ref NP_5</a>	(-) GNTVEAIAPKSPKK

- 1. [gi|113582|sp|P14639|](#)
- 2. [gi|18599827|ref|XP\\_0](#)
- 3. [gi|20890288|ref|XP\\_1](#)
- 4. [gi|22128700|gb|AAM92](#)
- 5. [gi|20890144|ref|XP\\_1](#)
- 6. [gi|4140371|gb|AAD037](#)
- 7. [gi|10954995|ref|NP\\_0](#)

# Mascot results

Peptide Summary Report (A few peptides from an LCMS run) - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites Media

Links Customize Links MBlastV3 gpf MFC Ask a Question SASMail Google MegaBlast SEQUEST Browser TurboSequest Debug

Address [http://www.matrixscience.com/cgi/master\\_results.pl?file=../data/F981123.dat](http://www.matrixscience.com/cgi/master_results.pl?file=../data/F981123.dat) Go

## Mascot Search Results

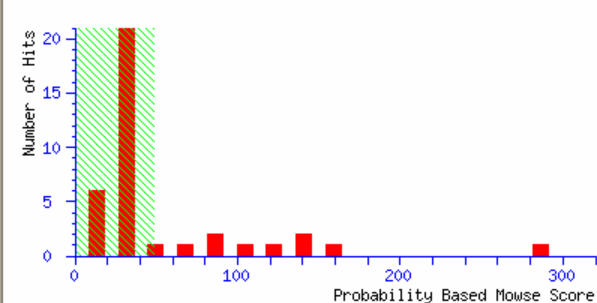
User : Cat R. Piller  
Email : [crp@brassica.com](mailto:crp@brassica.com)  
Search title : A few peptides from an LCMS run  
MS data file : U:\Mascot test data\TSQ\dyckall\_1.asc  
Database : MSDB 20020121 (820227 sequences; 255696937 residues)  
Timestamp : 24 Feb 2002 at 16:39:46 GMT

Significant hits:

- [Q9XZJ2](#) HEAT SHOCK PROTEIN 70.- *Crassostrea gigas* (Pacific oyster).
- [S14992](#) dnaK-type molecular chaperone hsp70 - soybean
- [A36333](#) dnaK-type molecular chaperone Hsc70-4 - fruit fly (*Drosophila melanogaster*)
- [A03309](#) dnaK-type molecular chaperone Dsim/Hsc70-1 - fruit fly (*Drosophila simulans*) (fragments)
- [Q95PU3](#) HEAT SHOCK PROTEIN (HSP70).- *Euplotes crassus*.

### Probability Based Mowse Score

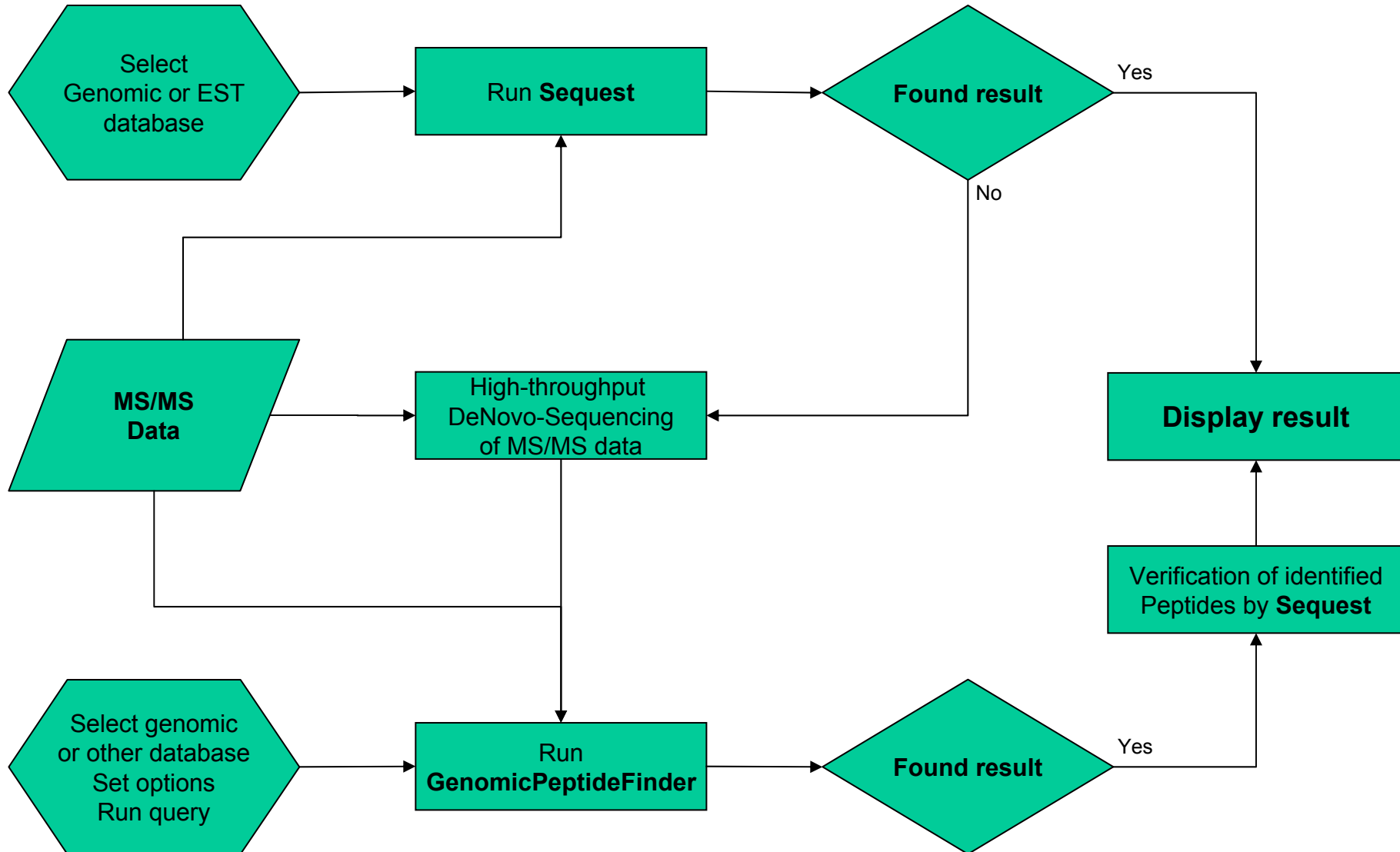
Ions score is  $-10 \cdot \log(P)$ , where P is the probability that the observed match is a random event.  
Individual ions scores  $> 48$  indicate identity or extensive homology ( $p < 0.05$ ).  
Protein scores are derived from ions scores as a non-probabilistic basis for ranking protein hits.



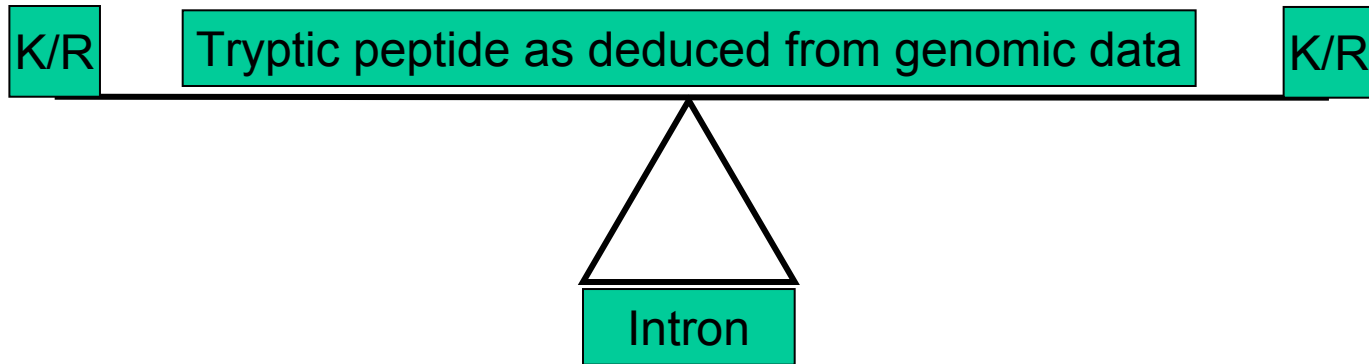
# Sequest and Mascot

- Sequest
  - Building of artificial spectrum
  - significance of XCorr
    - >1.5 for singly charged ions
    - >2.5 for doubly charged ions
    - >3.5 for triply charged ions
- Mascot
  - Probability model details not published
  - Score given as  $-10\log(P)$
  - Significance threshold is  $P = 0.05$

# GenomicPeptideFinder



# The problem with introns



20 to 25% of all detected peptides are split by introns on the genomic level and are thus not identifiable.

# Validation of genomic data

*Lhca p15.1* nuclear gene as an example

>[scaffold\\_3122](#) (Lhca2/4, p15.1)

MAALMQKSALS SRPACSTRSSRRRAVVVRAAADRK**LWAPGVVAPEYKGDLAGDYGWDPLGLGADPT**  
**ALKW**-*intron*

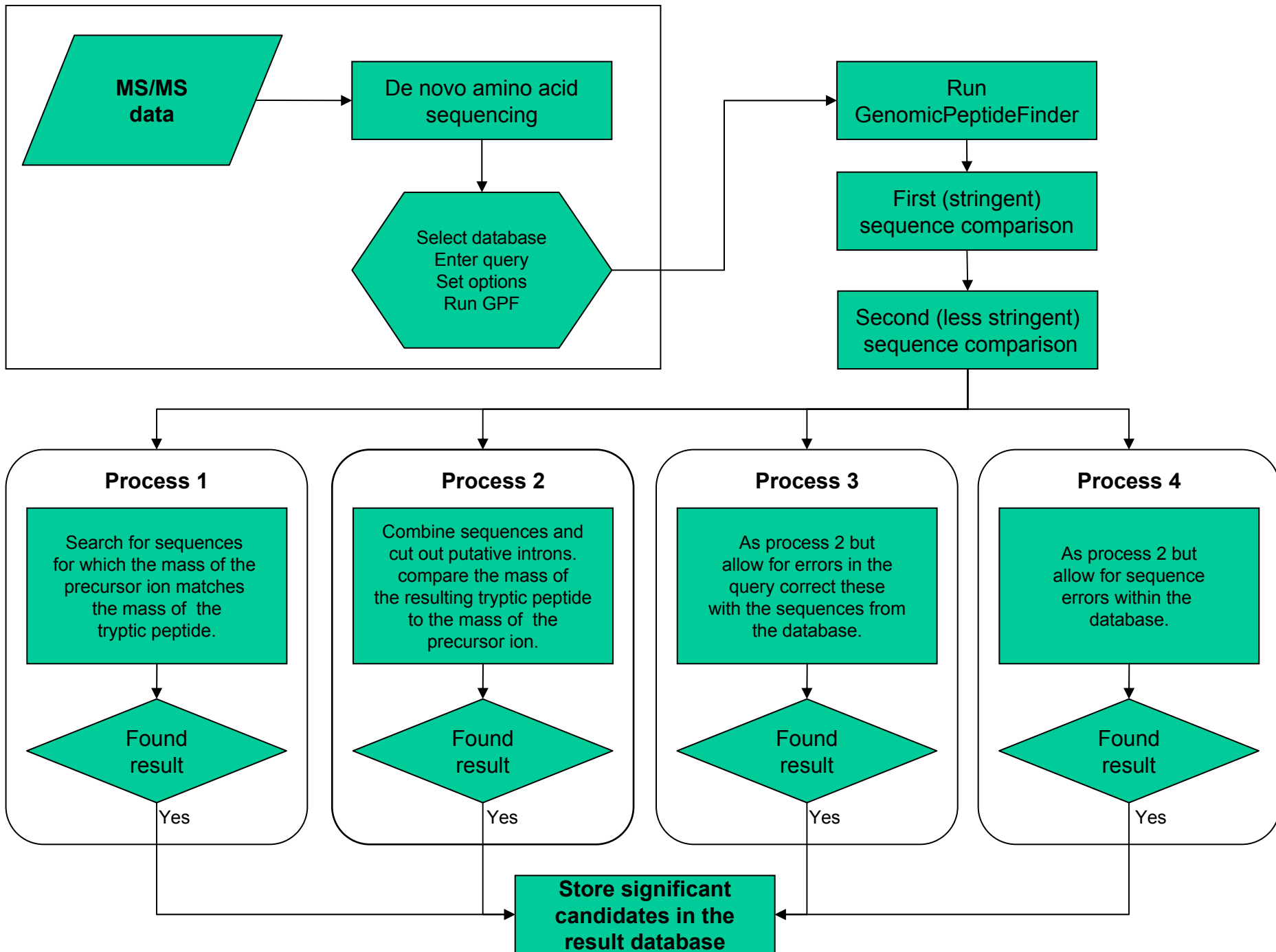
YRQSELQHARWAMLGVAGVLVQEIVKPDVYFYEAGLPQNLPEPFTNINMGLLAWEFILMHWVEV  
RRWQDYK**NFGSVNE**-*intron*

**DPIFK**<sup>GNK</sup>**VPNPEMGYPGGIFDPFG**-*intron*

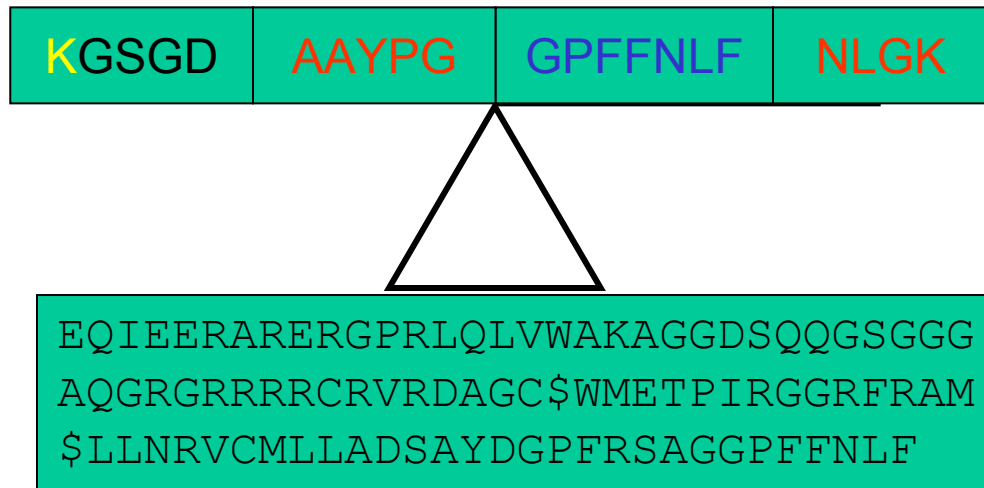
**FSK**GNLKEIQTKEIKNGRLAM-*intron*

IAYMAFILQAQATGKGPLAALSAHLSNPFGNILK-*intron*

NIGTCTVPHSVDVQGLTIPLTCLWPGSQ



# Example (Process 3)



**Query:** [RZ] AAYPG [VV] CFNPYNLKG | 2027.95

**Result:** GSGDAAYPGGPFFNLFNLGK

# Problems addressable by GPF

- Low abundant Proteins
- No EST or protein sequence available
- Introns
- False annotation
- Posttranslational processing

# Useful tools in mass spectrometry

CLASP, Resing et al 2004, Anal. Chem., 76, 3556-3568

OLAV, Collinge et al 2003, Proteomics, 3, 1454-1463

CHOMPER, Eddes et al 2002, Proteomics, 2 1097-1103

RADARS, Field et al 2002, Proteomics, 2, 36-47

Unnamed, Anderson et al 2002, J. Prot. Res., 2, 137-146

TurboGenomics, Cheung et al, Graphically-Enabled Integration of Bioinformatics  
Tools Allowing Parallel Executions

# RADARS

- Rapid automated data archiving and retrieval software
- Compatible with most data formats
- Consistent processing
- Automation
- Relational database
- New statistics

# OLAV

- Identifies peptides in a database from their spectra.
- Score based on signal detection theory
- Also exploits mass spectrometric information extensively
- Introduces structural matching
- Separates true from false positives

# CHOMPER

- Rapid evaluation of search results
  - Based on Sequest
- Uses .dta, .out and .html files generated by Sequest
- Parameters evaluated are
  - Charge state, MassA, DelMass
  - Xcorr, DelCn,Sp,RSp,Ions
  - Peptide sequence
- Score dependent on thresholds set for the combination of the above Parameters

# CLASP

- Score
  - Chemical properties
  - Integrates Sequest and Mascot
  - Quality of the spectra
- 19-29% false positives in Sequest and Mascot
- Comparable to CHOMPER

# Anderson et al

- Unnamed software comparable to CLASP and CHOMPER
- Support vector machine learning algorithm
- Using all reported Sequest scores and four more parameters
  - Number of peaks and fraction matched
  - Total ion current
  - Sequence similarity between top and consecutive matches

# TurboGenomics

- Software to integrate bioinformatic tools
- Graphical programming
- CGI compatible
- Problem
  - Most available programs are not compatible to common object request broker (COBRA) standard

# De novo sequencing

- Assigning amino acid sequences to a CID spectrum
- Mass difference in between two peaks can represent an amino acid
- Ion series can thus define a peptide
- Problem with incomplete ion series
  - Non breaking of certain bonds
  - Noisy spectra
  - Starting point of series hard to determine

# Current approaches to de novo sequencing

- Exhaustive generation of all possible sequence combinations that could account for a precursor mass
- Defining subsequences and maximizing the score for non overlapping patches
- Graph theory: reduces all ions to y-ions, reading along the generated ion-series may reveal the sequence
- Binary trees: a node represents a peak the border to a neighbor is defined by the difference in mass being exactly the mass of one amino acid

# DeNovoX

nsi78\_11.1803.1806.2.dta  
Sequence

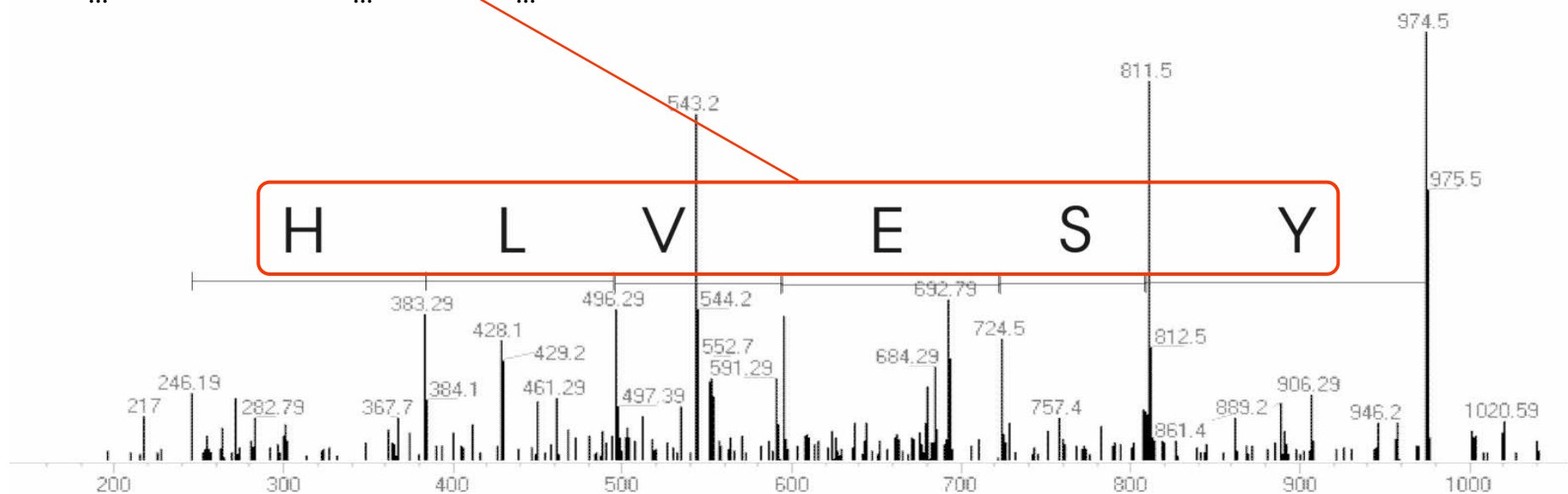
[LW]QYSEVLH[AR]

[PD]SQYSQVLH[AR]

WLQYSEVLH[AR]

...

Absolute Probability	Relative Probability
6.4%	41%
11.5%	31.6%
4.3%	7.2%
...	...



# De novo sequencing programs

	Graph theory	User similar	Dynamic programming	Bayesian
DeNovoX		+		
PEAKS	+		+	
Lutefisk	+			
SeqMS	+			
Sherenga	+			
MassSeq				+

Overview, Standing, KG, Curr. Op. Struct. Biol., 13, 595-6011

PEAKS, Ma et al 2003, Rapid Com. Mass Spec., 17, 2337-2342

Lutefisk, Johnson et al 1997, Rapid Com. Mass Spec., 11, 1067-1075

# Theory graph approach

- Converting all possible ions to b-ions or y-ions

N-terminal ions	Conversion to corresponding b-ion
a	Ion+28
a-NH3	Ion+45
a-H2O	Ion+46
b	Ion
b-NH3	Ion+17
b-H2O	Ions+18
C-Terminal ions	
y	Precursor-Ion+2
y-NH3	Precursor-Ion-15
y-H2O	Precursor-Ion-16

# PEAKS

- Removes noise and centers peaks
- Deconvolutes doubly and triply charged ions to singly charge state
- Building a database of the 10000 best suitable amino acid combinations for the given precursor mass
- Selection of the prediction with the largest number of high abundance peaks

$$score = f\left(\frac{h_1}{h}\right) \times f\left(\frac{h_2}{h}\right) \times f\left(\frac{h_3}{h}\right) \times e^{-\left(\frac{m'-m}{\delta}\right)^2} \times \log h$$

f(x) = function evaluating the presence of an ion and its supporting ions

h = abundance of the peak ( $h_x = x$ -ion abundance)

m' = observed mass

m = theoretical mass

delta = mass error

# Lutefisk

- Theory graph approach
  - Transforming all ions to b-ions
- Finding subsequences and then joining them
- Finding highest scoring assembly
- Rescoring best results with cross correlation, fast Fourier transformation and ion-count
- Returning a weighted summation of the above scores

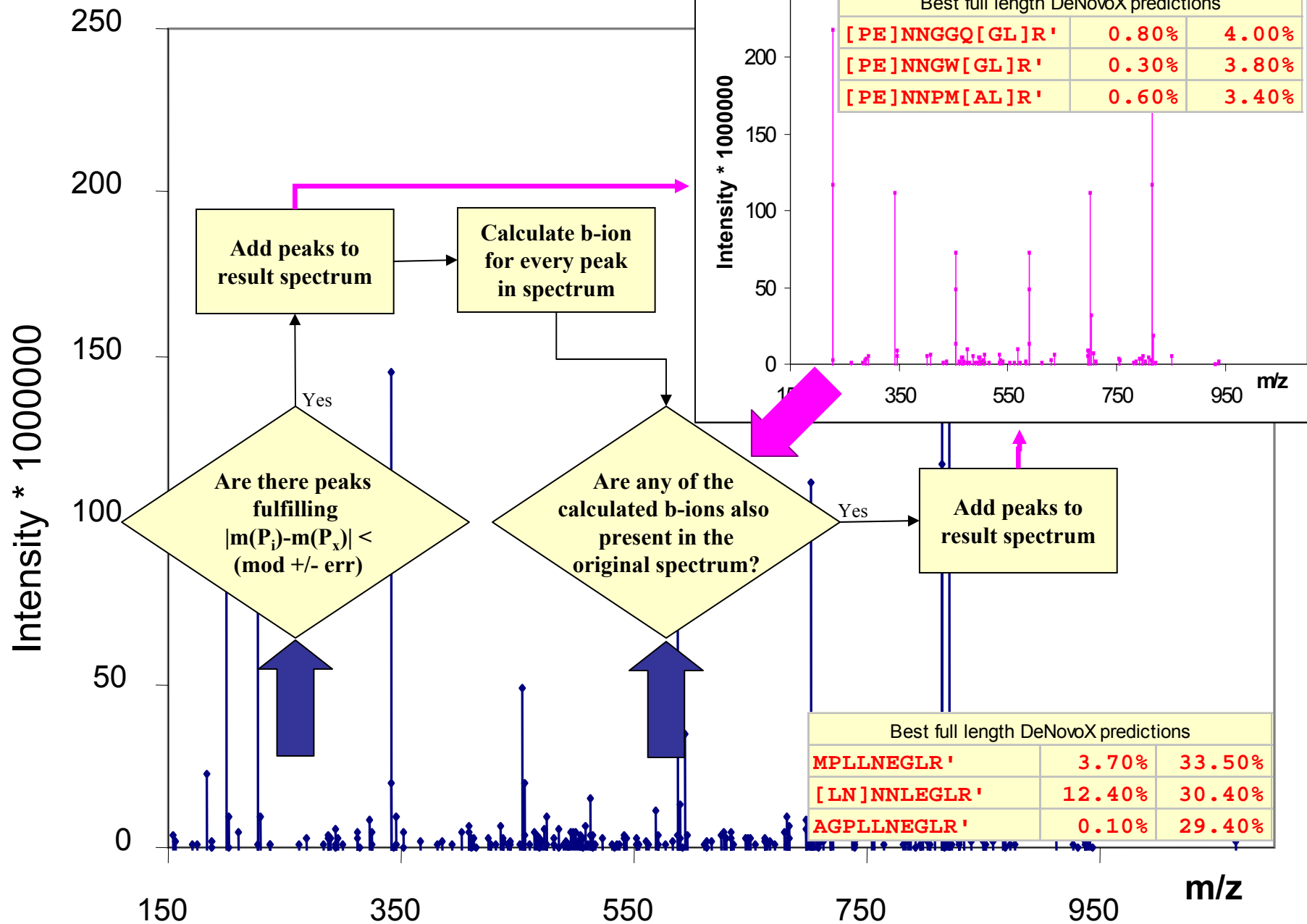
# Possible improvements

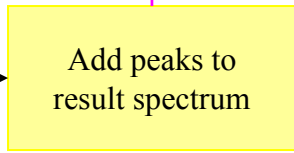
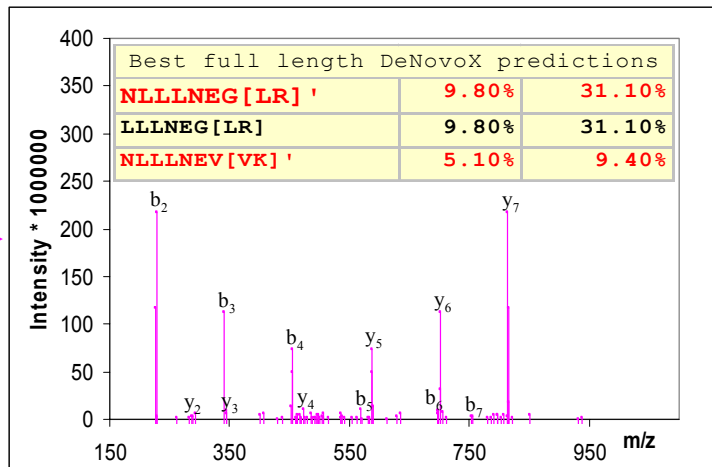
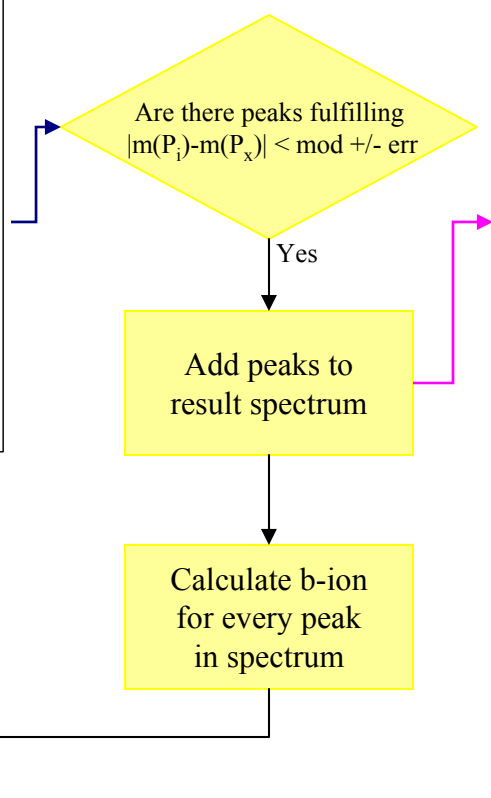
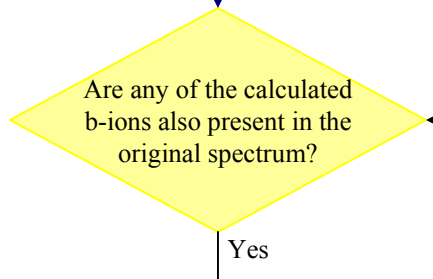
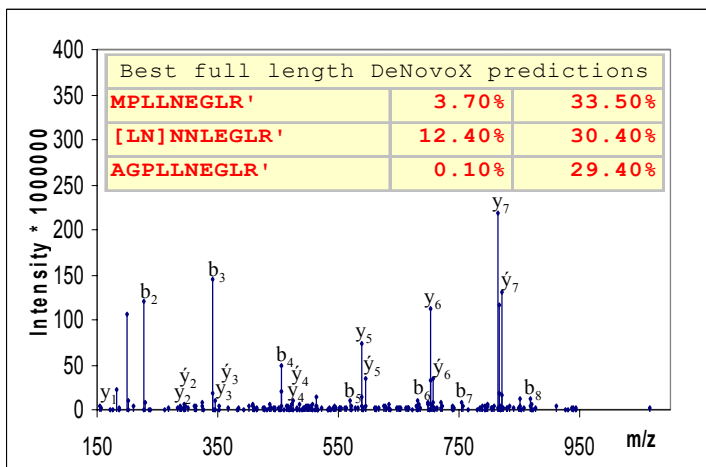
- Differential labeling of peptides
- Extraction of  $\gamma$ -ions from spectra
- Resulting in clean spectra
- De novo sequencing of these spectra

Priska et al 2003, Mol. Cell. Prot., 2.7, 426-427

Cannon and Jarman 2003, Rapid Com. Mass Spec., 17, 1793-1801

Uttenweiler-Joseph et al 2001, Proteomics, 1, 668-682





Some ions resulting from collision induced fragmentation are shown in the spectrum. Y and b-ions, the most abundant ion types, are specially labeled with  $y_n$  and  $b_n$ , where n stands for the ion number index. Arginine labeled y-ions are depicted as  $\underline{y}_n$ . All carbon atoms are exchanged for their heavier isotope thus resulting in a 6 Dalton mass shift for a labeled Arginine.

When adding a b-ion to the result spectrum its intensity is set to the same intensity as the y-ion it is calculated from.

The result spectrum (78 peaks) contains visibly less ions than the original spectrum (217 peaks). Some Y-ions were lost due to the fact that the heavy y-ion was not present. The prediction for the peptide using DeNovoX (Thermo Finnigan) was significantly more accurate using the pre processed spectrum than using the spectrum containing both labeled and unlabeled ions.