Deriving statistical models for predicting peptide tandem MS product ion intensities

Frédéric Schütz & Terry Speed

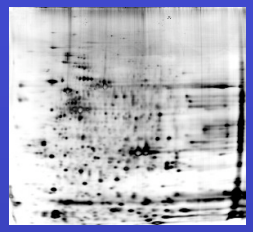
Division of Genetics & Bioinformatics
The Walter and Eliza Hall Institute of Medical Research

In collaboration with the Joint ProteomicS Laboratory
(WEHI/Ludwig Institute)

Introduction

- Proteomics is critical to our understanding of cellular biological processes
- Mass Spectrometry (MS) has emerged as a key platform in proteomics for the high-throughput identification of proteins
- Sophisticated algorithms, such as Mascot or Sequest, have been developed for database searching of tandem MS (MS/MS) data
- Major bottleneck: results must often be manually validated
- More robust algorithms are needed before the identification of MS/MS data can be fully automated

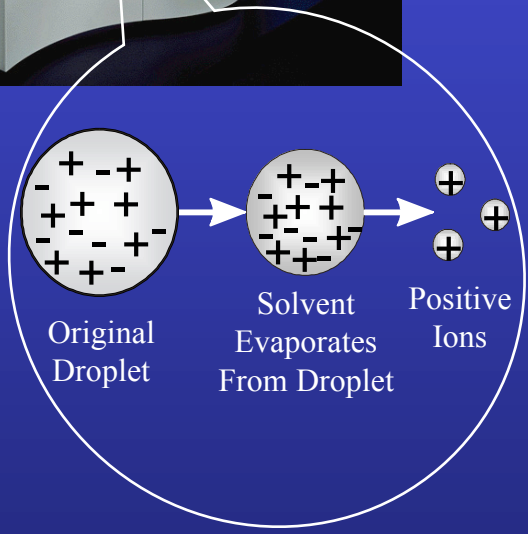
Tandem MS for protein identification



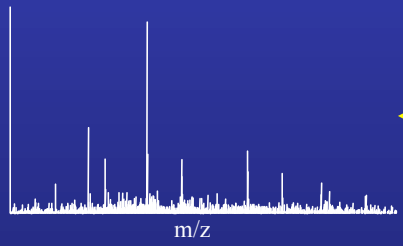
2-D Gel
(or 1-D Gel)

In-gel
Digest
(Trypsin)

Capillary
Column
RP-HPLC
(On-line;
60min Gradient)

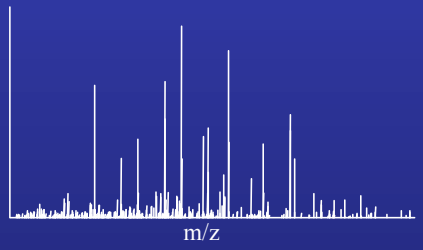


MS Analysis
(ESI Ion Trap)



MS data

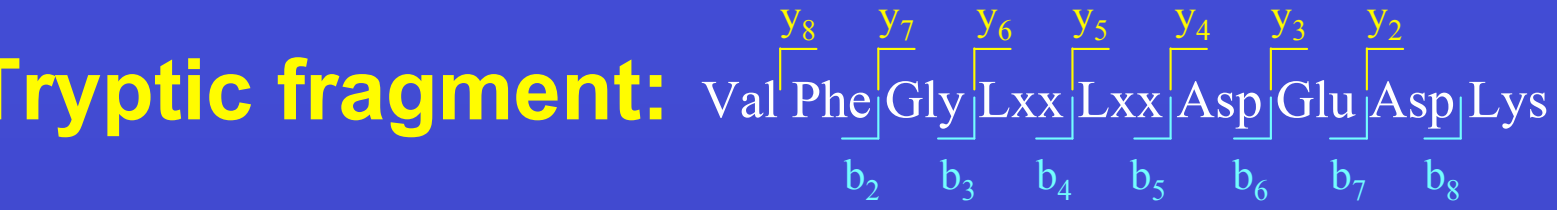
CID
(Most
intense
ion)

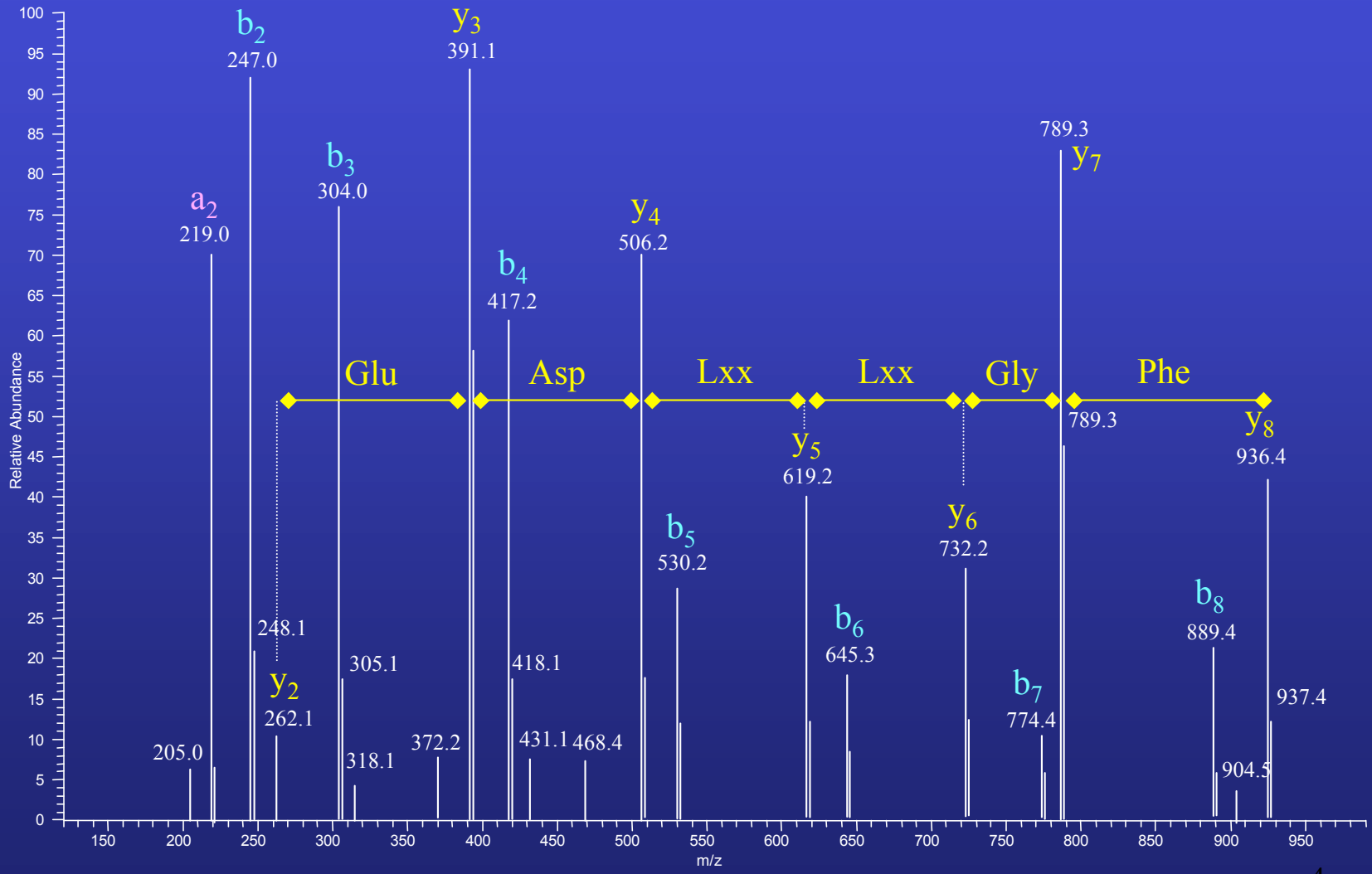


MS/MS data

CID = collision induced dissociation

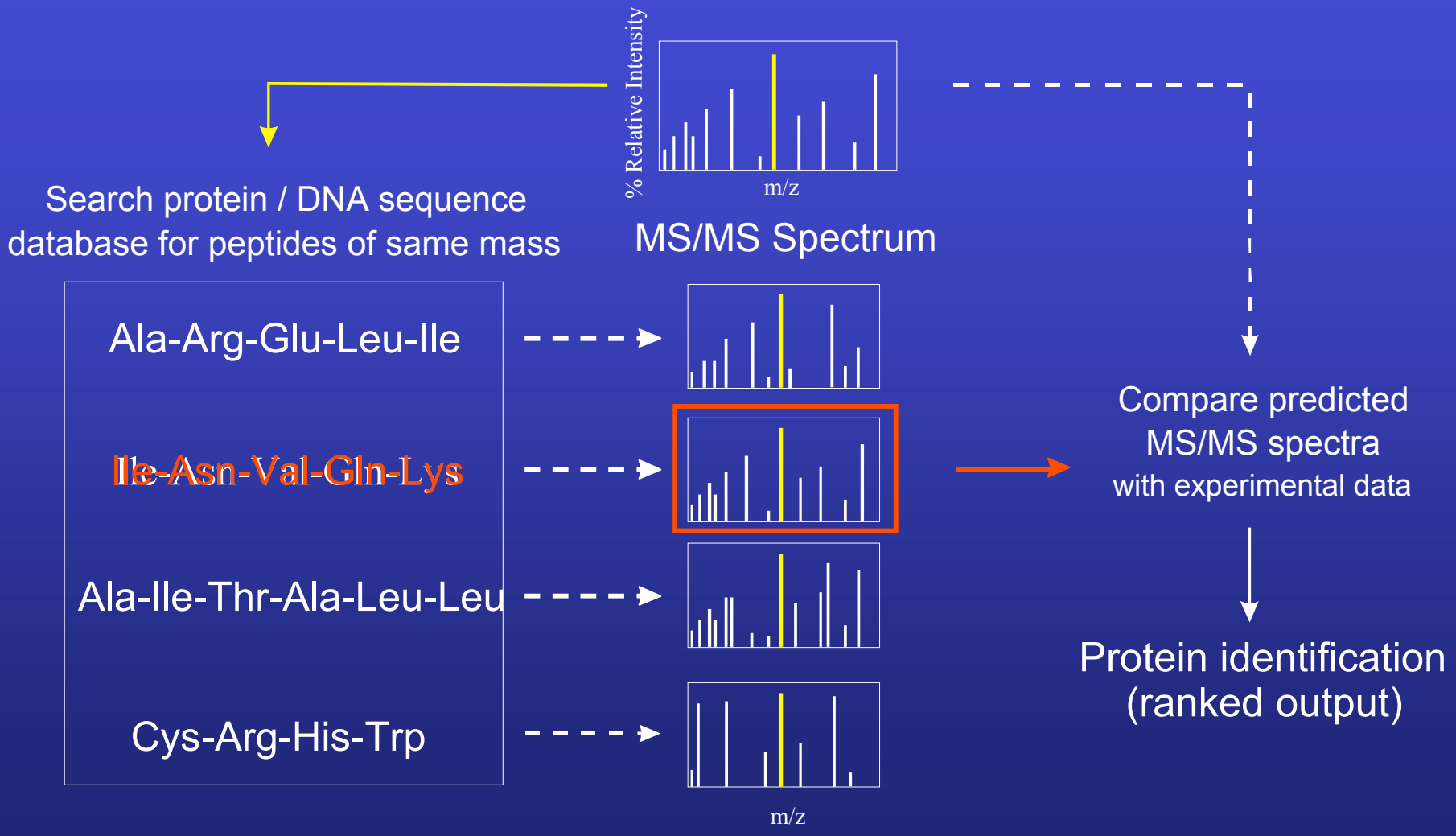
Tryptic fragment:





Example MS/MS spectrum

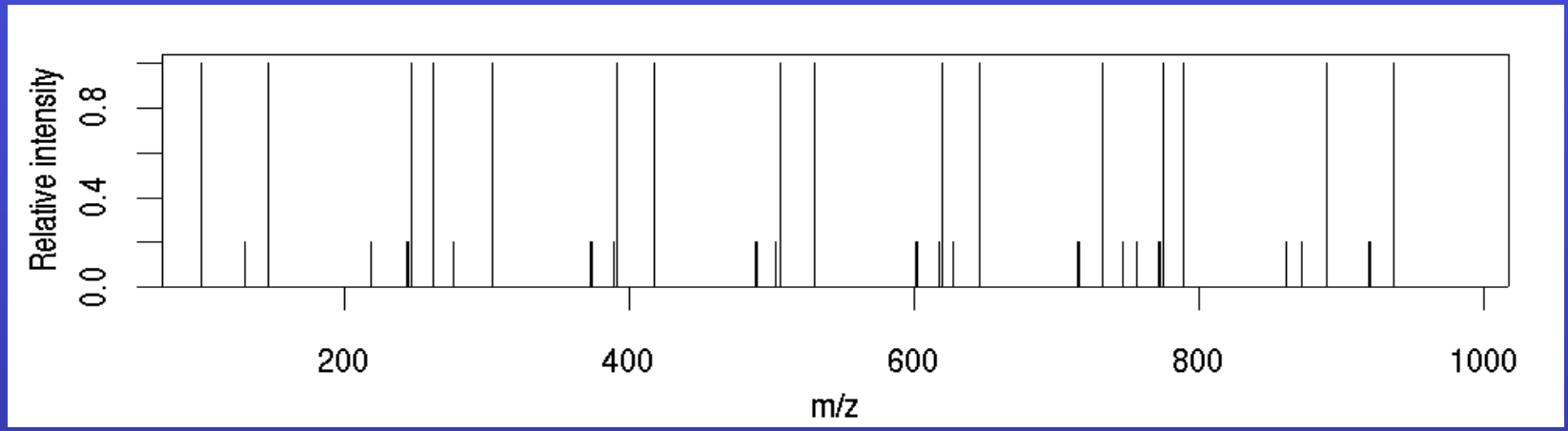
Database searching of MS/MS data



Comparing a sequence and a spectrum

- In order to do a comparison, you need to be able to predict a spectrum for a given peptide sequence
- Predicting the *position* of the peaks is not too hard: the masses of the different amino acids are known (barring any chemical modification) and you just have to add them using chemistry rules
- Most search algorithms do just this, and assume equal fragmentation for the different bonds (i.e. peaks of equal intensity, at least for the main ions)
  - SEQUEST: creates such a simplified spectrum, and correlates it with the experimental spectrum
  - MASCOT: use a probabilistic model based on the position of the peaks whose details are not public

Comparing a sequence and a spectrum, cont



Example of a spectrum predicted by Sequest

- This approach works well for identifying most peptides
- Several peptides exhibit fragment ions that differ greatly from this simple model
- Those peptides often yield low or insignificant scores, thus preventing a positive identification

Peptides do not fragment equally !

- Experimenters have known for a long time that some bonds are more likely to break than others
- Some peaks may appear only under certain conditions (e.g. some neutral losses require the presence of given amino acids)
- This knowledge is used to manually validate results that are returned by the search algorithms
- Examples:
  - Cleavage at Xaa-Pro (nP) are more frequent
  - Cleavages at Asp-Xaa (cD) are more frequent under certain circumstances

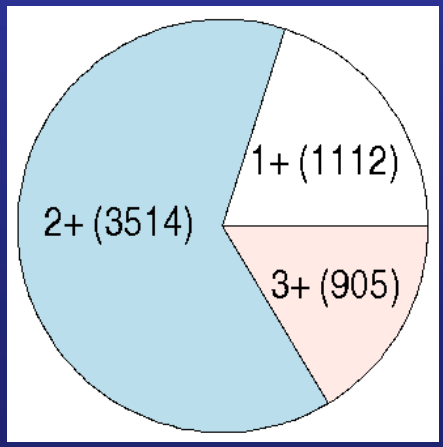
An example of ‘unusual’ spectrum

Top result from SEQUEST : **KPADPDNLK**, score 1.4148

(correct sequence, as found with another search algorithm and manual validation, is not in the top 10 scoring peptides returned by SEQUEST)

Can we predict the ion intensities ?

- If we don't have prior knowledge of chemical mechanisms which could increase or decrease the fragmentation of certain bonds, we can take results from previous experiments and try to find indication about such mechanisms.
- Our data, collected by the JPSL in Melbourne over several years:
  - 20,642 spectra from an ESI – Ion-Trap mass spectrometer
  - Database search: SEQUEST, using a non-redundant database of 1.4M proteins proteins provided by the Ludwig Institute in Lausanne
  - 12,540 spectra were positively identified after a manual validation**
- 5,531 unique sequences



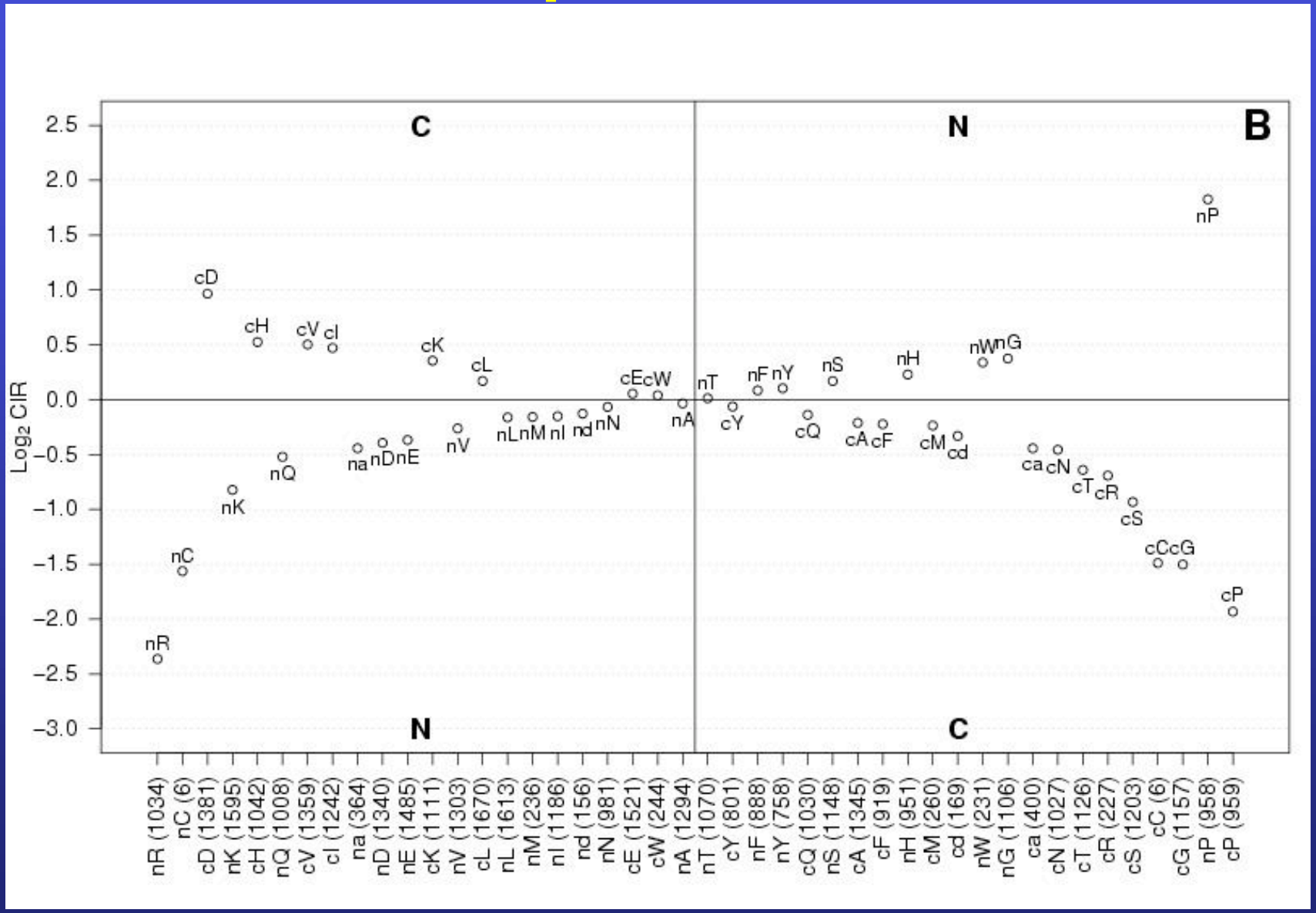
First method: Cleavage Intensity Ratios

- Method similar to the one used by different groups, with small differences
  - Huang *et al,* IJMS 2002, Breci *et al*., Anal. Chem. 2003
- The peak corresponding to known ions (b, y) are extracted from each spectrum
- The intensities of all the ions that were produced by the breakage of a given bond (b, y, several charges) are summed and compared to the average on all bonds (*s* = a given cleavage in a given peptide with N cleavages).

$$CIR_s = \frac{\sum_{z=1}^{Z} (b_s^{z+} + y_s^{z+})}{\frac{1}{N} \sum_{i=1}^{N} \sum_{z=1}^{Z} (b_i^{z+} + y_i^{z+})}$$

CIR	< 1	= 1	> 1
Cleavage	Reduced	Average	Enhanced

Example of results



Those CIR values are for the sum of all the ions for one particular cleavage, and not for one peak, and thus can not be directly used for peak prediction

‘Relative Proton Mobility’ Scale

If number of Arg residues ≥ number of charges	***Non-mobile***
If number of Arg, Lys & His < number of charges	***Mobile***
otherwise they are designated	***Partially-mobile***

Quantifying the Asp-Xaa (cD) bond cleavage

	Mobile	Partially-Mobile	Non-Mobile
1+	-	K_1 2.37 (238)	R_1 5.10 (126)
2+	K_1 0.81 (358) R_1 1.04 (316)	K_2 1.66 (276) K_1R_1 2.06 (301)	R_2 4.96 (92)
3+	H_1K_1 0.88 (54) K_1R_1 0.91 (37) R_2 1.31 (24)	$H_1K_1R_1$ 1.63 (79) K_1R_2 2.51 (21) $H_1K_2R_1$ 1.94 (23) $H_1K_1R_2$ 2.71 (10)	R_3 3.63 (12)

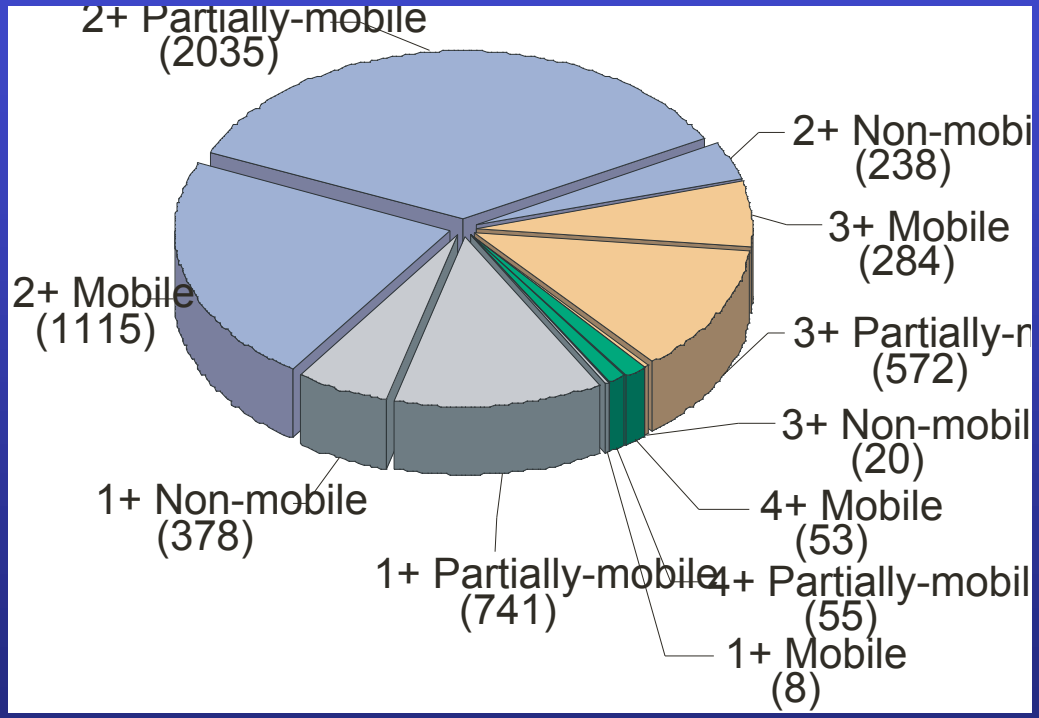
Entries: average CIR (#peptides), stratified by #basic residues

Limitations of the CIR approach

- While it has proved to be useful, the CIR approach has some limitations:
  - We can not assess the significance of each factor
  - For a given cleavage, the intensity is blindly credited to the 2 neighbouring amino acids, instead of being split
  - It is difficult to consider other variables that may have an influence on the fragmentation model

Second method: linear model

- Data were categorized into 9 different strata, according to charge state (1, 2 or 3+) & 'relative proton mobility' scale



- For each spectrum the peaks corresponding to the different ions (b, y) are extracted and normalised by the sum of intensities of identified peaks for this spectrum.

Second method: linear model, cont

- The following model is fitted:

$$\log_2(\text{normalised intensity of ion } j) = \mu + a_{c(j)} + b_{n(j)} + c \log_2(\text{peptide length}) + \text{error}$$

- where
  - μ is the baseline cleavage intensity (i.e. the average cleavage intensity if no factor has any special effect on fragmentation)
  - $a_{c(j)}$ is the increase/decrease of fragmentation due to *c(j),* the residue on the C-terminus of the bond
  - $b_{n(j)}$ is the increase/decrease of fragmentation due to *n(j),* the residue on the N-terminus of the bond
  - $\log_2(\text{peptide length})$ accounts for the lower intensity, due to the normalisation process, of a given cleavage when it occurs in a longer peptide

Variable selection

- The *a* and *b* variables are two sets of dummy variables (20 variables each for the 20 amino acids)
- For each of them, the factor that is the closest to the weigthed average intensity is removed from the model.
- In other words, one of the residues on each ‘side’ (C or N) is selected as the reference, the residue that ‘does nothing’ (parameterization using ‘treatment’ contrasts)
- Once a reference has been selected, backward selection is performed to remove all variables that are not significantly different from the reference at the 1% level.

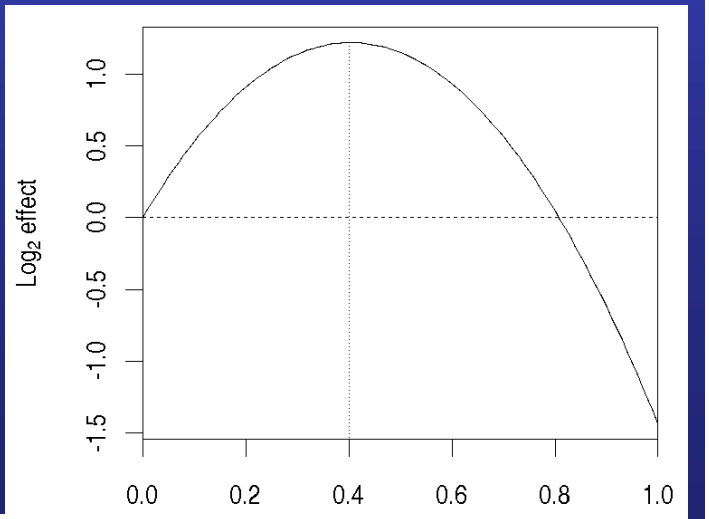
Adding other variables

- Peptides seem to fragment towards the middle rather than at their ends
- To account for this fact (also noticed by other groups), positional terms were added to the model:

$$log(rel\ int_j) = \mu + a_{c(j)} + b_{n(j)} + c\ log(peptide\ length) + \alpha(pos) + \beta(pos^2) + error$$

where *pos* (0<*pos*<1) is the relative position of the cleavage inside the peptide

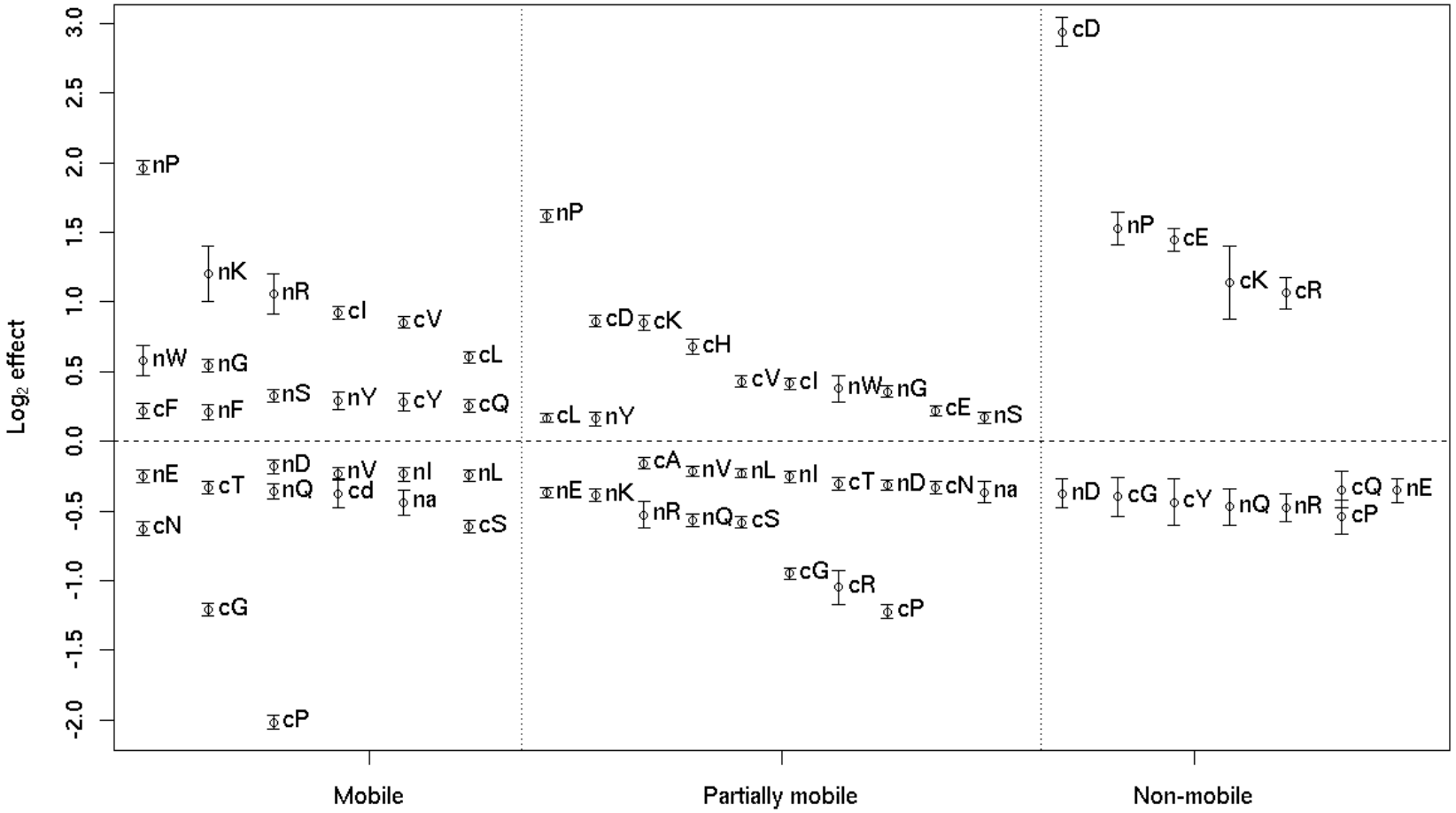
- The positional terms were retained in almost all of the models, and confirm the observations.



Log-effect of position for y ions, 2+, partially mobile

Example of results

Effects for doubly-charged precursor (y-) ions

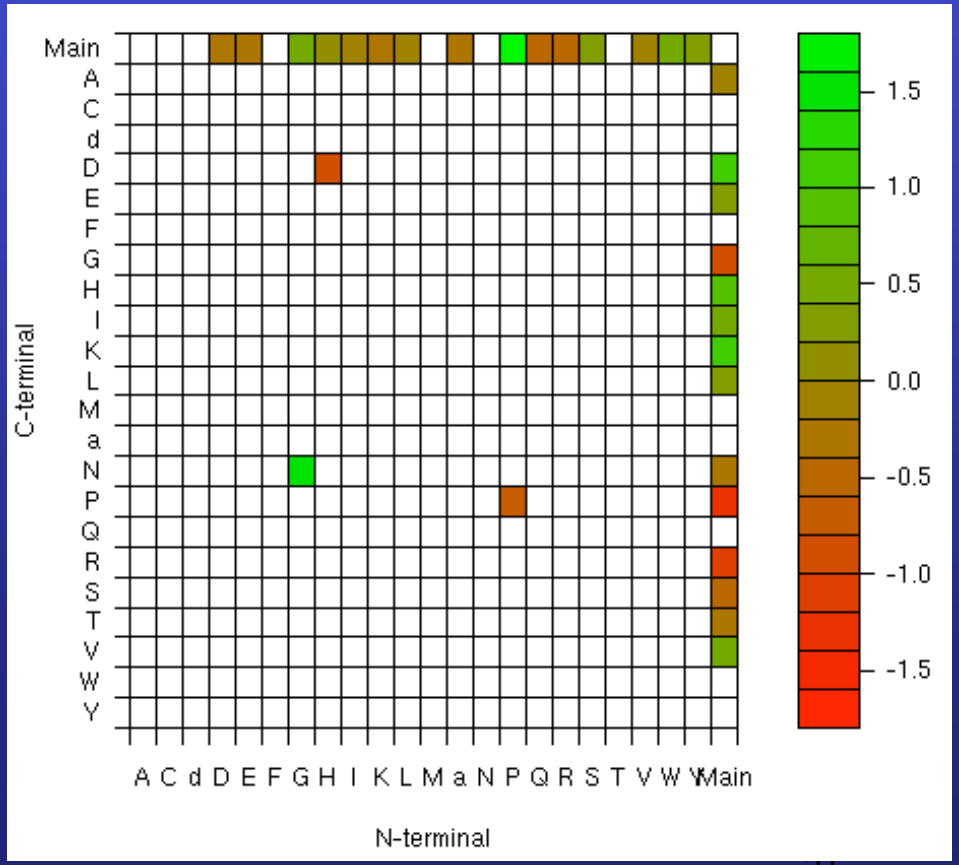


Interactions

- There are 20x20=400 possible interactions between the different pairs of amino acids (without including potential modifications as separate amino acids)
- We can not add them all to the model and select the ones that are significant
- Strategy:
  - We fit a simple linear model including only the positional variables
  - We extract the residuals and arrange them in a table (rows are labelled according to the C-terminal amino acid, columns by the N-terminal amino acid)
  - We apply the median polish algorithm to extract effects for the rows and columns. The residuals are then an indication of the intensity of the interactions
  - The interactions are entered into the model in decreasing order of absolute magnitude of the residuals and selected using the forward algorithm.

Interactions: results

- Very few interactions have any significant effect across more than one stratum
- Main exception: cleavage of NG bonds (Asn-Gly) is enhanced in most of the cases



Log-effect of residues for y ions, 2+, partially mobile

Limitations of the linear model

- Several artefacts appear in the linear model:
  - The $\log_2$(peptide length) is not really a factor that influence fragmentation; it is mainly a artefact due to normalisation and (ideally) should not appear in our model
  - Indirect effects: if a factor has a very large effect (e.g. nP, enhancing), then the corresponding factor on the other side (cP) will seem to have a slight opposite effect
    - if a bond has a P on its C-terminal (Pro-Xaa), then there is almost certainly a cleavage Yaa-Pro somewhere else in the peptide, meaning that there is a large peak in the spectrum, and by normalisation, the peaks corresponding to Pro-Xaa will look smaller even if cP does not have any direct effect on the cleavage.

Third method: mixed or random effects linear model

- With the linear model, all data points (peak intensities) are considered to be independent.
- But this is not an accurate description: different peaks that come from the same spectrum share something (they have been normalised together) and this should be taken into account.
- We group the data and fit a mixed model

$$\log_2(\text{relative intensity of ion } j \text{ in peptide } i) = \mu + a_{c(j)} + b_{n(j)} + c \log_2(\text{peptide length}) + \alpha(\text{pos}) + \beta(\text{pos}^2) + g_i + \text{error}$$

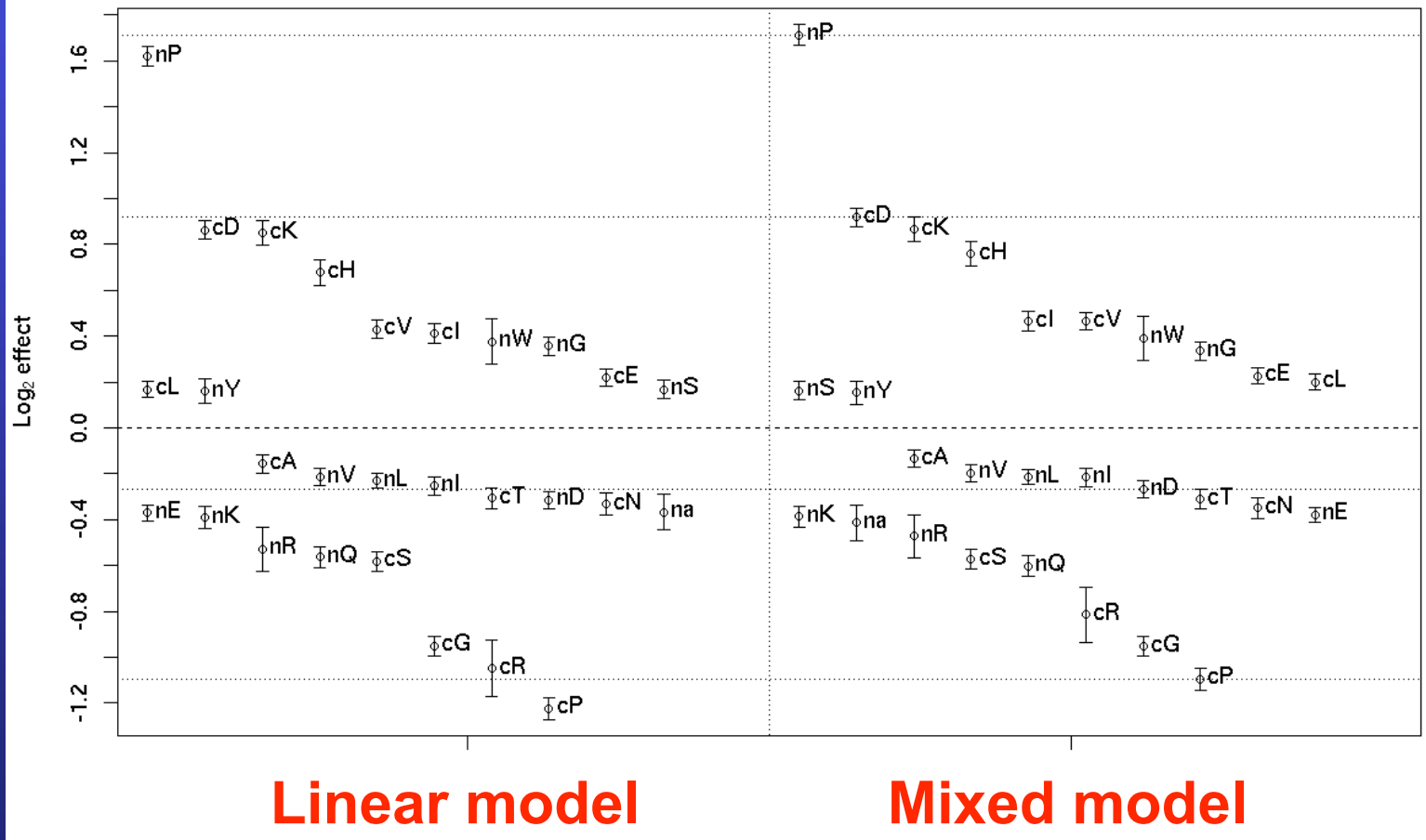
where $g_i \sim N(0, \sigma^2_g)$ are random peptide effects.

Mixed model: results

- As expected, the coefficient for the $log_2(length)$ is slightly lower in each strata (but still significantly different from 0)
- The other coefficients do not change enormously and the sets of selected variables is almost the same as in the linear model
- For effects that go in pairs (e.g. cD and nD) and have opposite effects (one positive and one negative), the changes are as expected, i.e. they both go in the same direction, the largest effect becoming larger and the other one getting closer to 0.

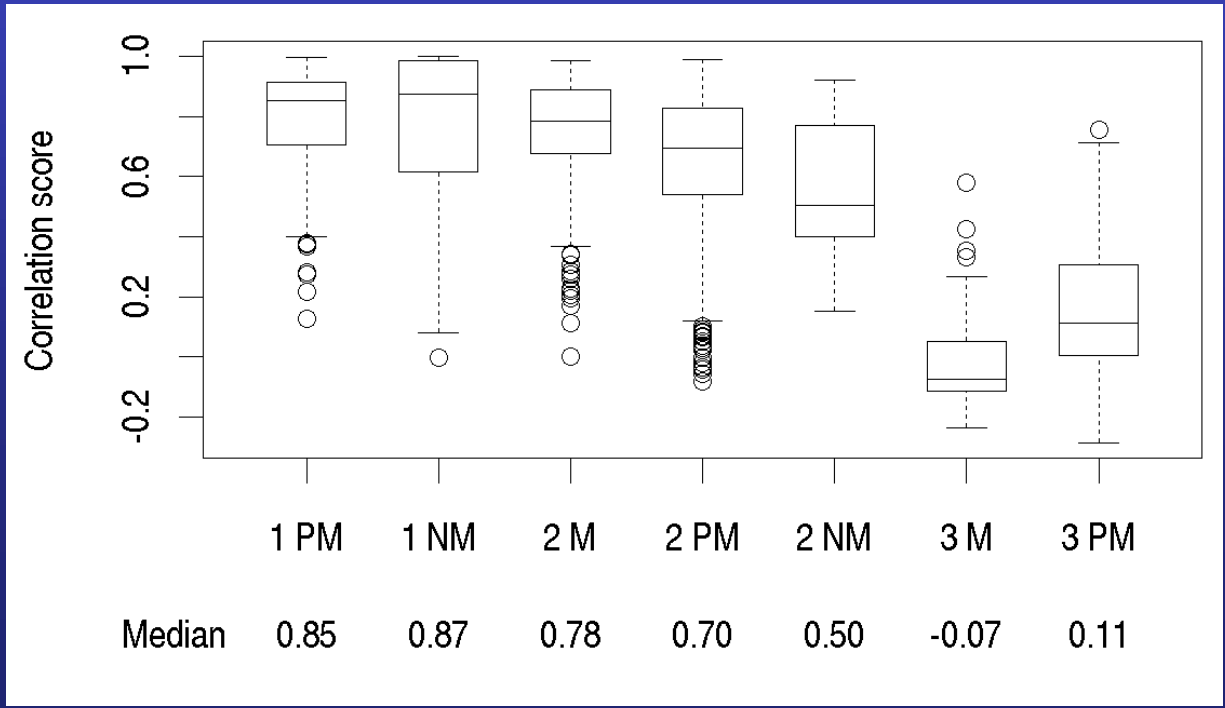
Mixed model results: some details

Effects for doubly charged precursor ions

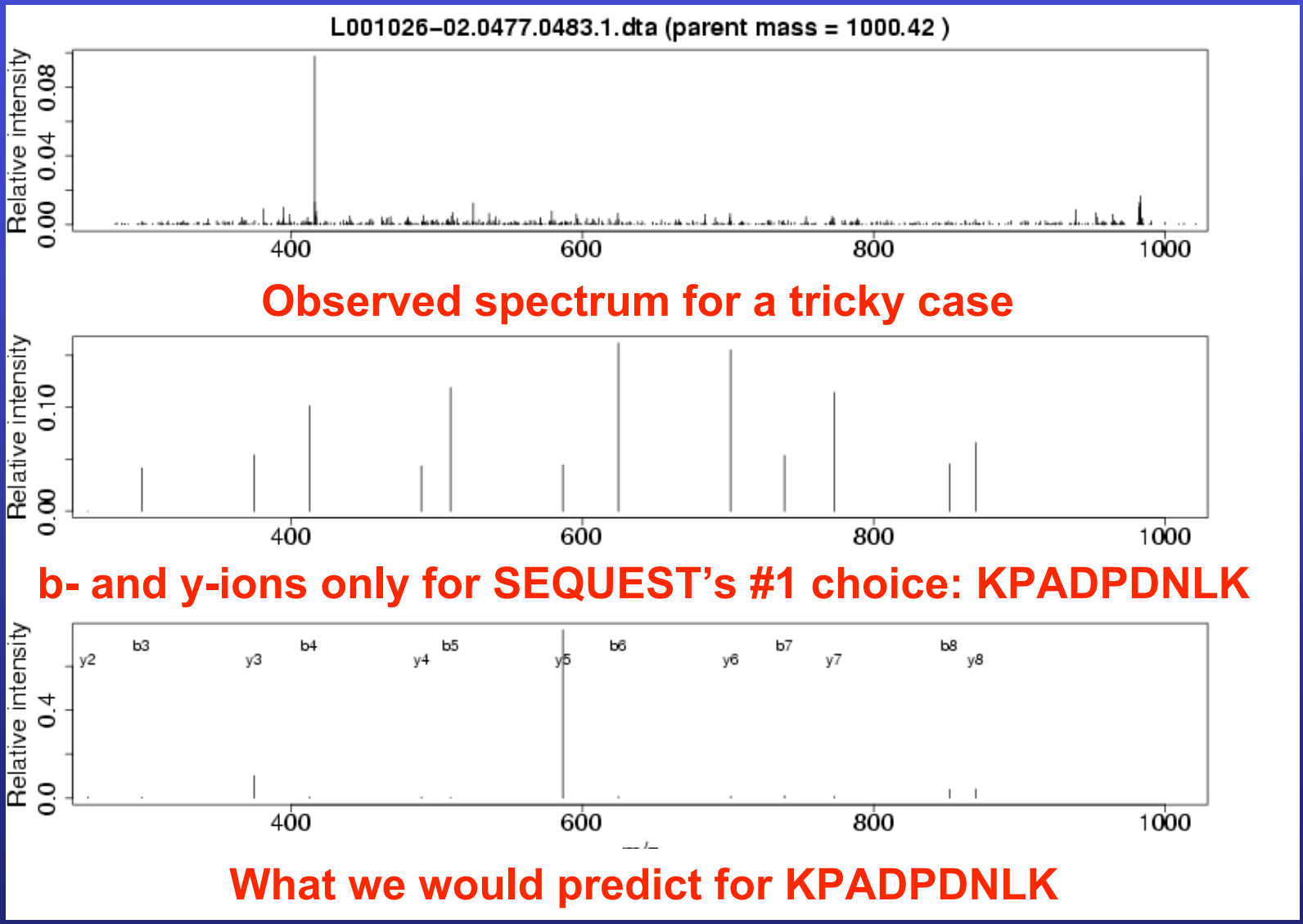


Prediction results

- The predictions were tested on a set of 1,254 peptides that were not used to fit the models
- Prediction of triply-charged spectra is currently not much better than random prediction (expected since the corresponding spectra are usually complex and noisy)
- Many of the spectra with low correlations are actually peptides that are **not tryptic** (e.g. `RAELEAK`, doubly charged, correlation -0.19) and for which it is expected that the fragmentation pattern will be different.

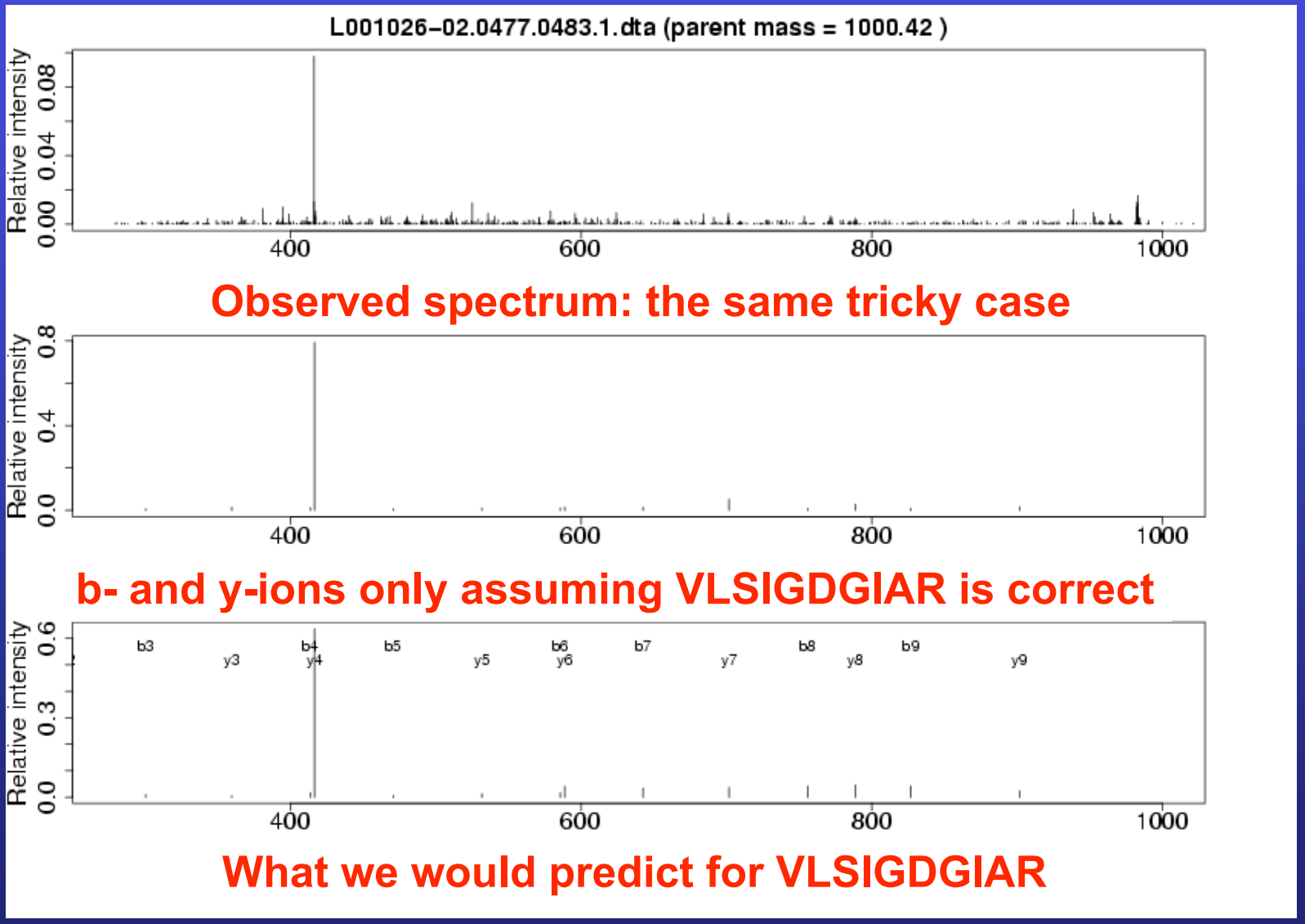


Our model for a SEQUEST #1 choice



Correlation coefficient between the bottom two spectra: -0.107

Correct answer: not in SEQUEST's top 10



Correlation coefficient between the bottom two spectra: 0.995

Conclusions and ongoing work

- Prediction of spectra is becoming feasible
- Better search algorithms are expected
- The current models are already useful for helping the validation of proteins identification (either automatically if the score is high enough or manually)
- Clearly, more fixed effects should be taken into account, including charge localisation (to better account for non-tryptic peptides)

- Our current dataset do not really allow us to test whether a new algorithm based on these models performs better than SEQUEST (since **all** our test cases have been identified with SEQUEST, we cannot do better !)
- We are currently building a new test set that should solve this problem

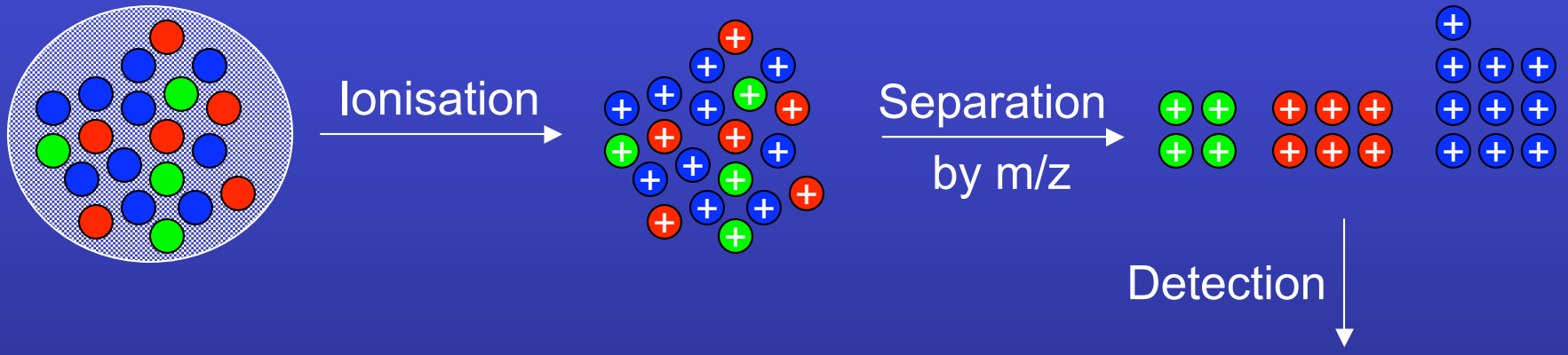
- These are important steps towards fully automated identification of peptide MS/MS data

Acknowledgments

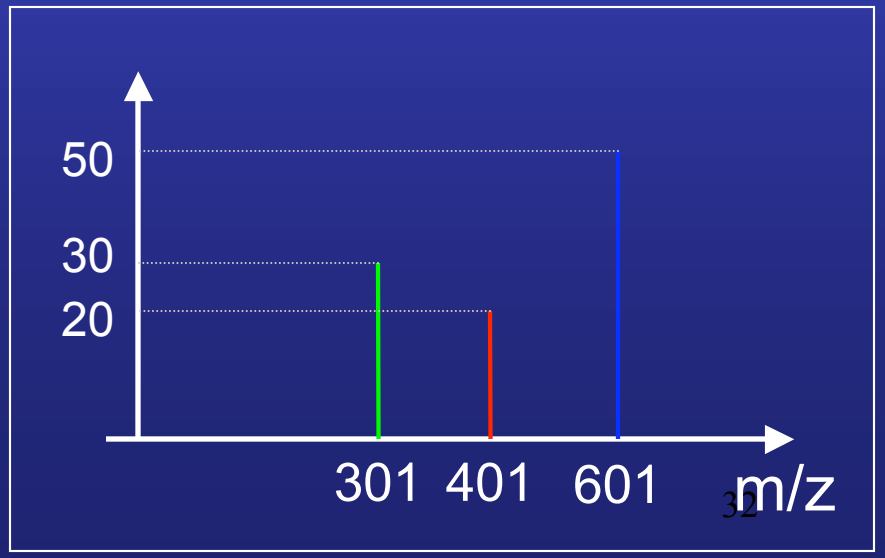
- **Joint ProteomicS Laboratory Ludwig Institute, Melbourne**
  - Eugene Kapp
  - James Eddes
  - Gavin Reid
  - Lisa Connolly
  - David Frecklington
  - Robert Moritz
  - Richard Simpson
- **Bioinformatics, WEHI**
  - Gordon Smyth
- **Dept. of Chemistry, Melbourne University**
  - Richard O 'Hair

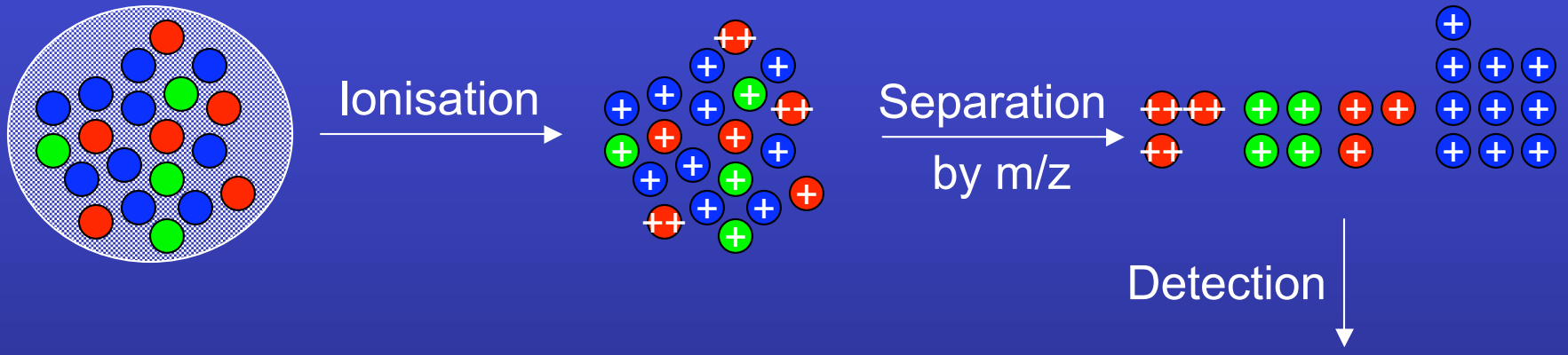
What is a Mass Spectrometer ?

"*An analytical device that* ***determines the molecular weight of chemical compounds*** *by separating molecular ions according to their mass-to-charge ratio (m/z)*"

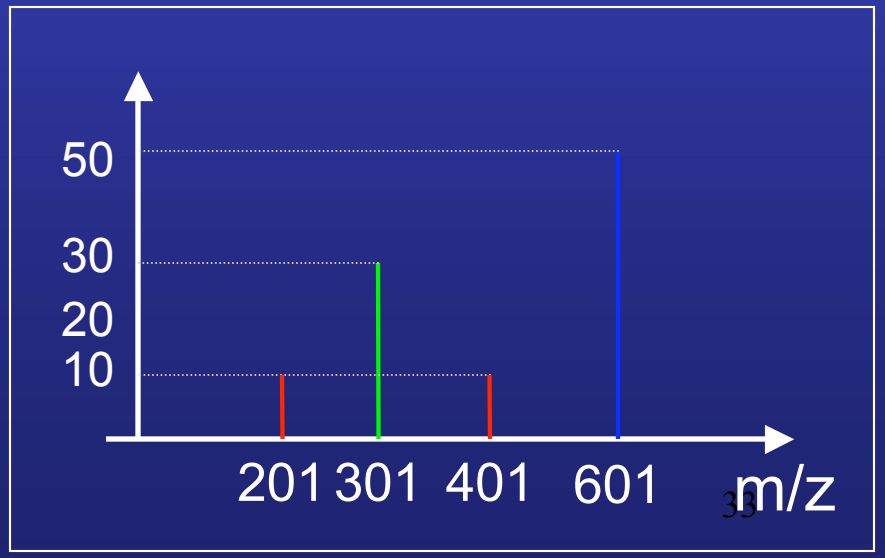


- molecular weight = 600 Da
  abundance = 50 %
- molecular weight = 400 Da
  abundance = 20 %
- molecular weight = 300 Da
  abundance = 30 %



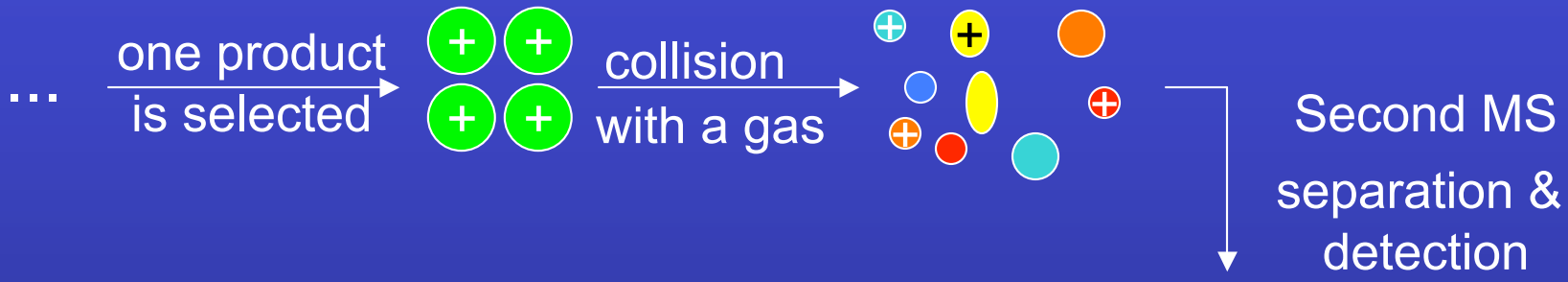


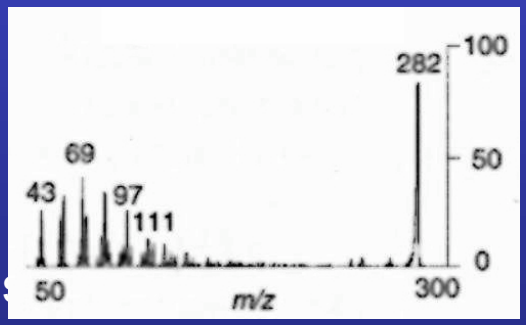
- molecular weight = 600 Da
  abundance = 50 %
- molecular weight = 400 Da
  abundance = 20 %
- molecular weight = 300 Da
  abundance = 30 %



Tandem MS (MS/MS)

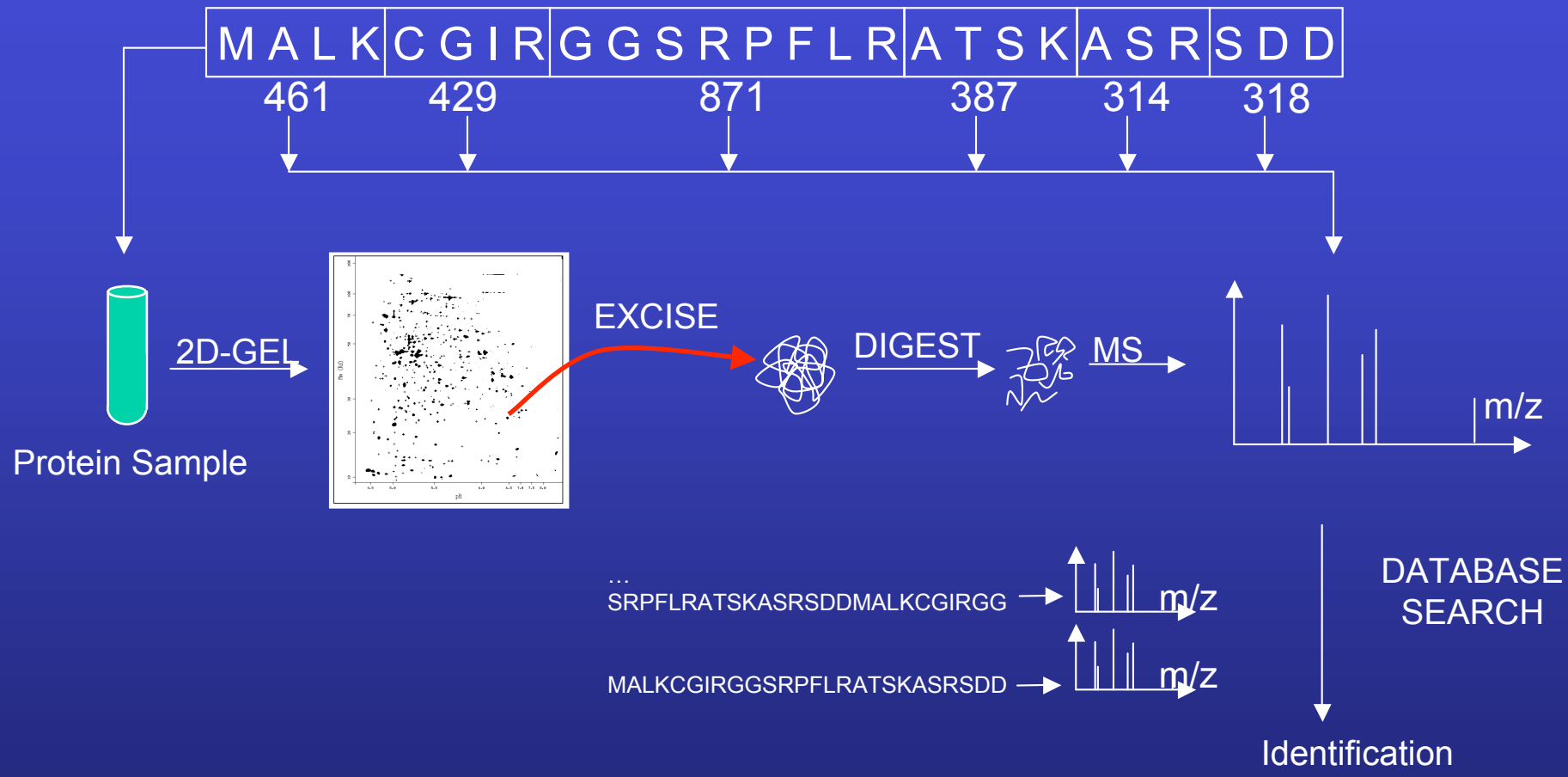
- To gain structural information about the detected masses:





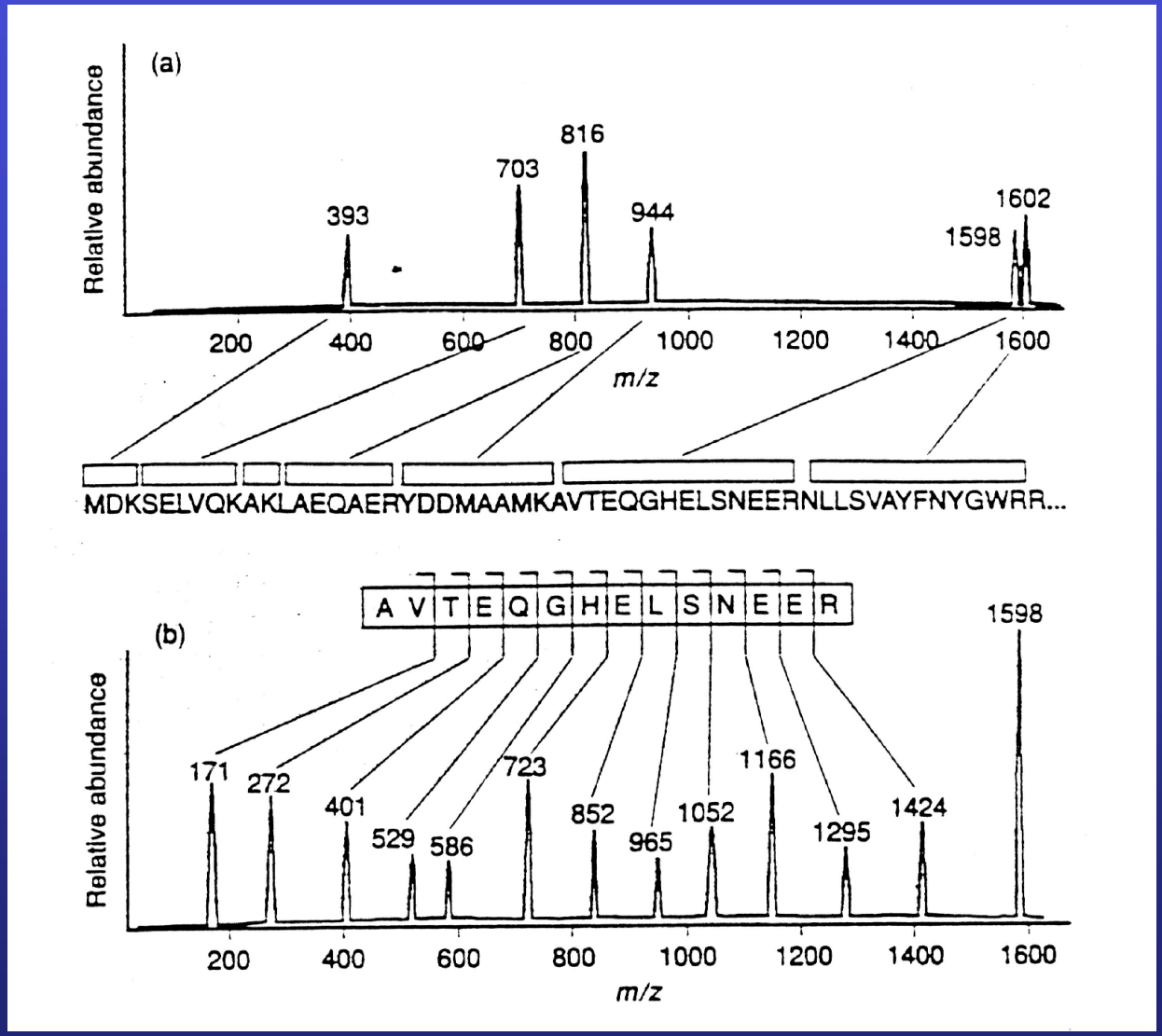
  - different molecules of the same substance can
  - in each molecule, only the pieces that retain one of the charge will be observed and present in the spectrum; the others are discarded.

Peptide Mass Fingerprinting



- The exact protein needs to be in the database
- Works only with simple protein mixtures

Comparison between PMF and Tandem MS



Direct Interpretation of MS/MS data

- « De Novo sequencing »
- Direct interpretation of the mass differences between peaks
- The only identification method available if the peptide is not in the database
- It can give useful information (partial sequence) for database search
- The spectrum must be of good quality (not too many peaks, main peaks high above noise level)

De Novo sequencing: EGVNDNEEGFFSAR