Proteómica

Agulhas em palheiros

Mass list to protein...

1. Protein cut in a predictable manner

   Sizes of the pieces - Fingerprint

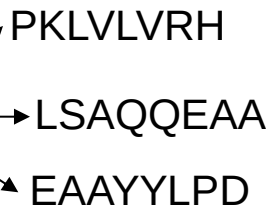
2. Protein sequence in a database cut *in silico* (in the same way) – Fingerprint of all proteins in database

Compare 1 with 2

How are the pieces generated ?

Using proteases that cut specifically in some amino acid residues

partial hydrolisys - PEPTIDES



Different protein sequences – different peptides (different fingerprint)

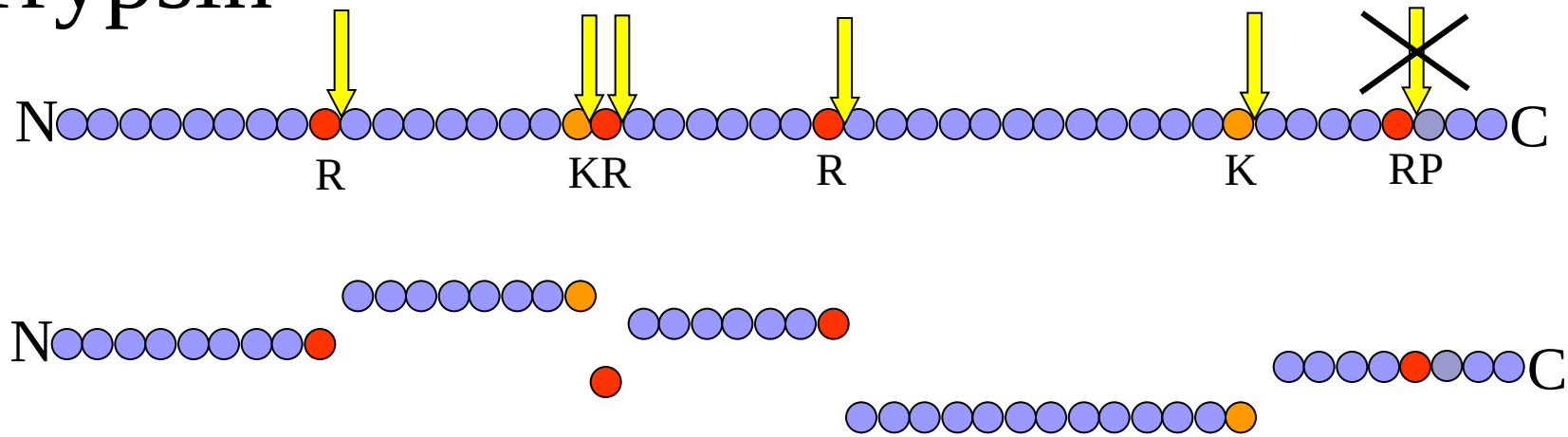
Not useful: Nonspecific proteases

Total hydrolisys – AMINO ACIDS

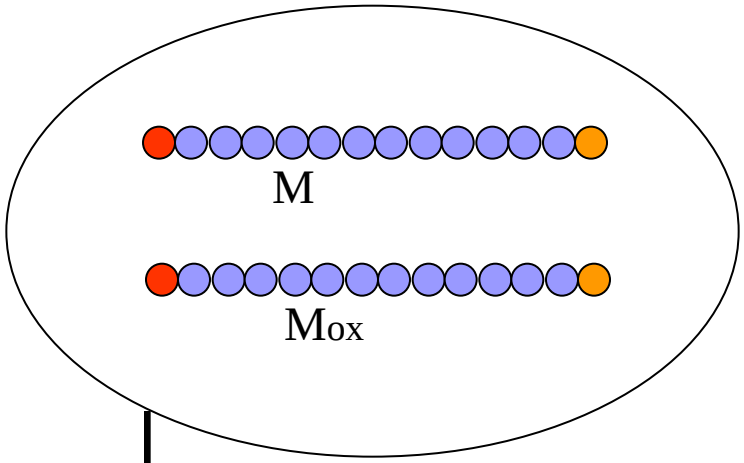
Name	Cleave	Don't cleave	N or C term
Trypsin	K, R	P	CTERM
Arg-C	R	P	CTERM
Asp-N	B, D		NTERM
Chymotrypsin	F,Y,W,L,I,V,M	P	CTERM
CNBr *	M		CTERM
Formic_acid *	D		CTERM
Lys-C	K	P	CTERM

* protease-like reaction mechanism

Trypsin



Missed cleavage sites



Mass change in peptides with methionine residues (16 Da)

Why scoring algorithms?

Is the right fingerprint or not...

True – if we have complete fingerprint

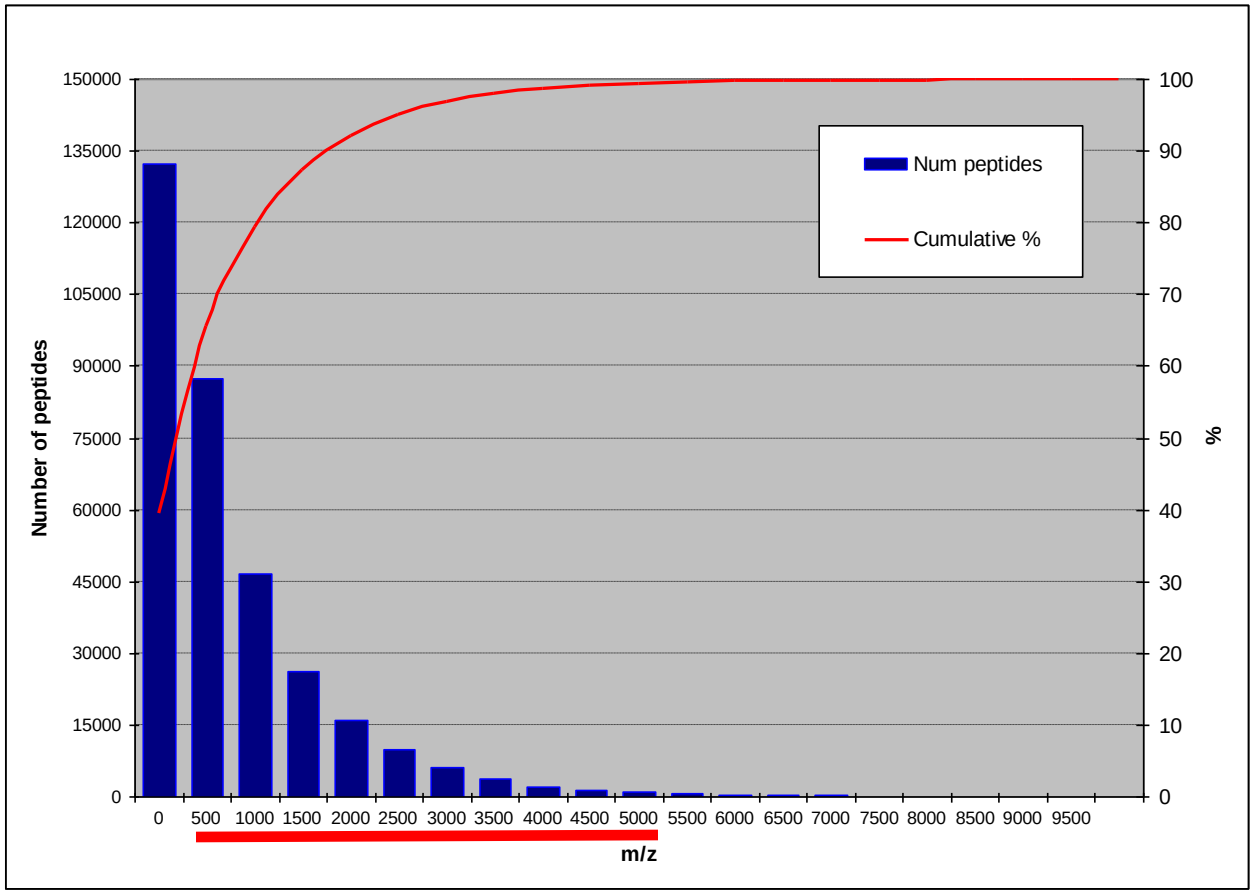
Protein: detect all peptides

(100% sequence coverage)

Never observed in real life

(laboratory)

Complexity?

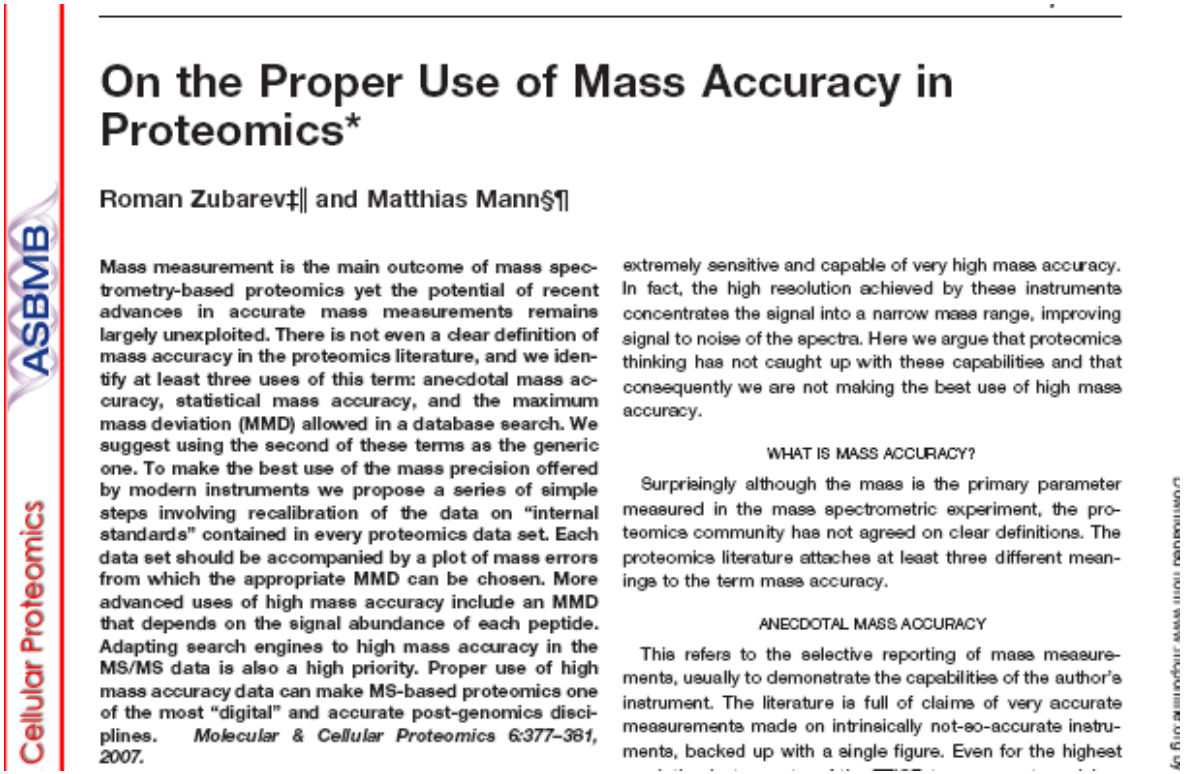


S. cerevisiae 5880 ORF – 150000 Péptidos (trypsin)

Mass accuracy

- Anedoctal
- Statistical
- Range

$$\frac{\Delta m}{m} \times 10^6 \; ppm$$



Cellular Proteomics

ASBMB

On the Proper Use of Mass Accuracy in Proteomics*

Roman Zubarev‡‖ and Matthias Mann§¶

Mass measurement is the main outcome of mass spectrometry-based proteomics yet the potential of recent advances in accurate mass measurements remains largely unexploited. There is not even a clear definition of mass accuracy in the proteomics literature, and we identify at least three uses of this term: anecdotal mass accuracy, statistical mass accuracy, and the maximum mass deviation (MMD) allowed in a database search. We suggest using the second of these terms as the generic one. To make the best use of the mass precision offered by modern instruments we propose a series of simple steps involving recalibration of the data on "internal standards" contained in every proteomics data set. Each data set should be accompanied by a plot of mass errors from which the appropriate MMD can be chosen. More advanced uses of high mass accuracy include an MMD that depends on the signal abundance of each peptide. Adapting search engines to high mass accuracy in the MS/MS data is also a high priority. Proper use of high mass accuracy data can make MS-based proteomics one of the most "digital" and accurate post-genomics disciplines. *Molecular & Cellular Proteomics 6:377–381, 2007.*

extremely sensitive and capable of very high mass accuracy. In fact, the high resolution achieved by these instruments concentrates the signal into a narrow mass range, improving signal to noise of the spectra. Here we argue that proteomics thinking has not caught up with these capabilities and that consequently we are not making the best use of high mass accuracy.

WHAT IS MASS ACCURACY?

Surprisingly although the mass is the primary parameter measured in the mass spectrometric experiment, the proteomics community has not agreed on clear definitions. The proteomics literature attaches at least three different meanings to the term mass accuracy.

ANECDOTAL MASS ACCURACY

This refers to the selective reporting of mass measurements, usually to demonstrate the capabilities of the author's instrument. The literature is full of claims of very accurate measurements made on intrinsically not-so-accurate instruments, backed up with a single figure. Even for the highest

The MOWSE
&
The MASCOT

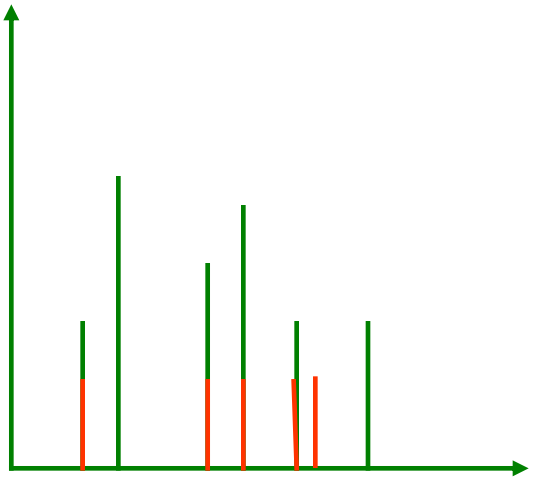
Perkins & Pappin 1999

Scoring methods

The simplest way :

Number of measured peptide masses corresponding to theoretical peptide masses

=> The proteins with the highest number of matching peptides are reported



- MOlecular Weigth SEarch (MOWSE) scoring algorithms (taking into account the relative abundance of the peptides in the database and the effect of the protein size)

- Probability scoring algorithms – MASCOT

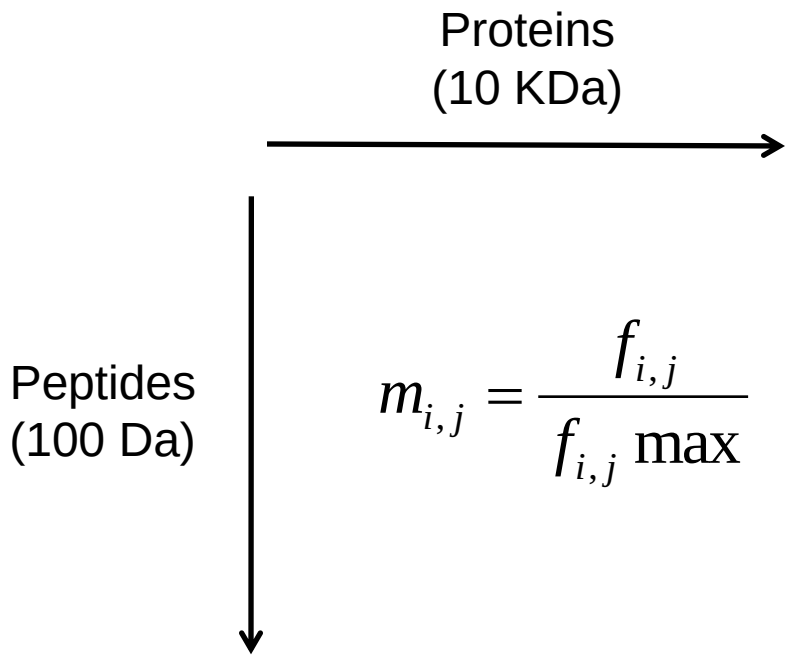
MOWSE

1. Compares calculated peptide mass and experimental data
2. Uses empirically determined factors
   - Statistical weight of each peptide match
   - Mowse factor elements
3. Score: (50 000)/Mprot x Π mi,j

Molecular weigth of the entry

Product from Mowse factor elements for each match peptide from the entry

Frequency factor matrix F



$$m_{i,j} = \frac{f_{i,j}}{f_{i,j}\ \max}$$

MOWSE

- Takes into account the relative abundance of peptides and the size of proteins

Disadvantages...

– The model consist of numerous spaces separated by 100 Da (per each 10 Kda)

– Does not provide a measure of confidence for the prediction

Why probability?

- Probability 6 of 10 peptides match a sequence is 10^{-5} → 10^{6} sequences
- Significance P<0.05 (1 of 20)
  - Database of 10^{6} sequences – $5x10^{-8}$
- Probability that the observed match is a chance event

Lowest = The Best → Significant??

Depends on the size of the database

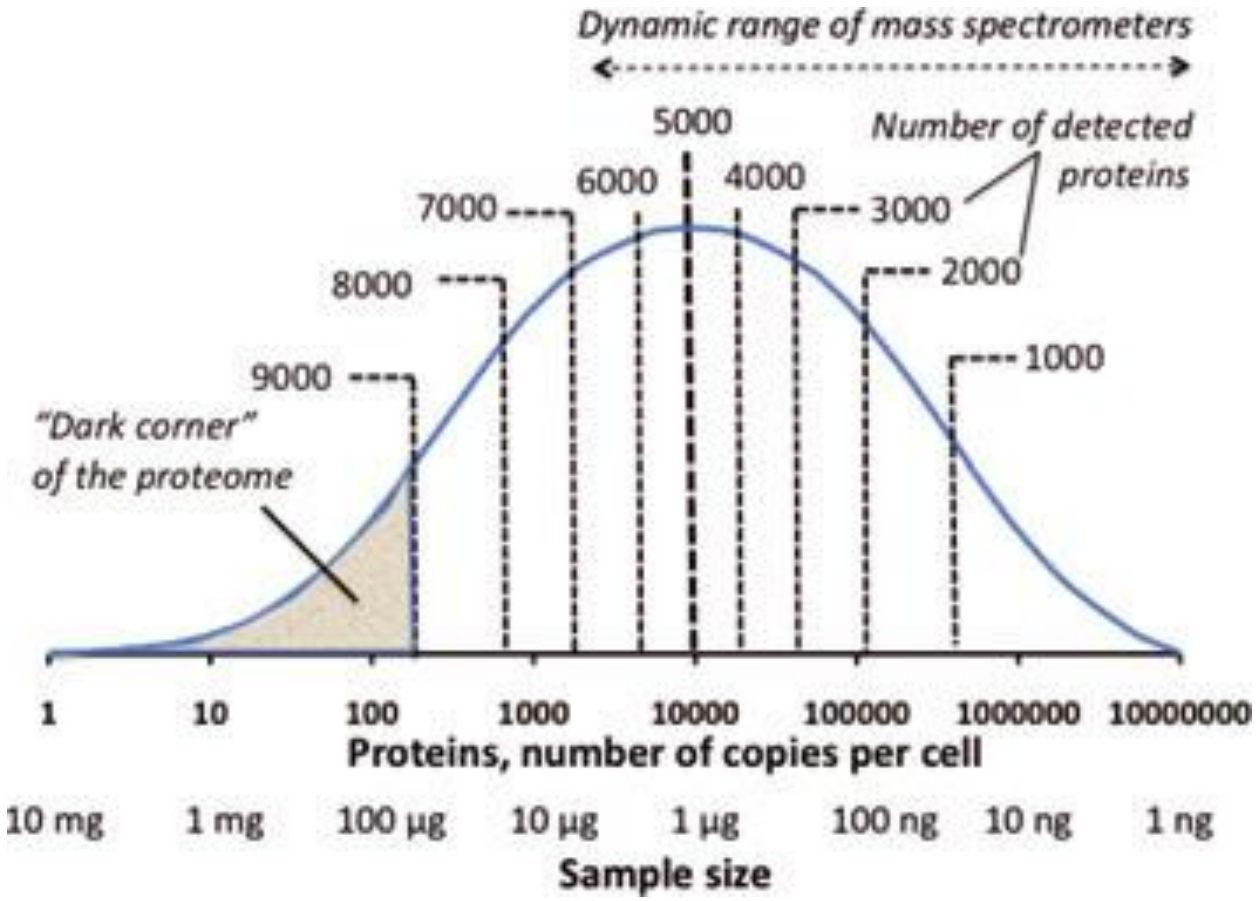
MASCOT

- Probability based MOWSE
- Probability that the observed match is a random event is calculated for each protein in the sequence database
- Mascot score: $S = -10\log(P)$
  - $P = 10^{-20}$ ► $S = 200$
- Probability model details not published

What influences the score ?

- Number of matched peptides
  - ↑ match - ↑ score
  - Depends on mass tolerance
- Number of calculated peptide masses to be matched
- Number of experimental peptides to be matched

The dark corner of the proteome



Other algorithms

- MaxQuant

- ProSight