PeakErazor – Calibrating MALDI Mass Spectra

Peter Højrup and Karin Hjernø

Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, Denmark.

PeakErazor.exe

1 - Introduction

Problem

Peptide mass searching/fingerprinting (PMS) is one of the most common methods for identifying proteins in proteomics. The method relies on identifying proteins based on the mass values of the peptides generated by enzymatic digestions, typically tryptic digestion. Very advanced search programs are available today, making identification of proteins by PMS using tryptic MALDI-TOF MS data quite straightforward.

Two criteria are essential for a successful identification: A high number of (significant) mass values and a high precision. In the current software we have tried to address both problems, with most emphasis on improving the precision.

In a given dataset the PMS can be improved by removing non-significant values, i.e. remove **peptide contamination**. 2-D gel separated proteins are typically contaminated by keratin and tryptic autodigest peptides, but other project specific contaminants may be present.

Including these peaks in the peptide mass search will obviously make the search less precise and consequently they have to be excluded from the peptide mass list before using the list as input in a protein identification.

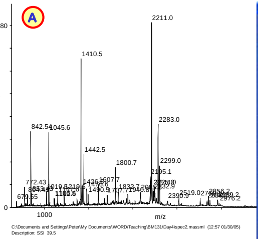
Calibration is usually performed as a two- or three-point calibration (either external or internal) of the mass spectra are used and in most cases the obtained accuracy is sufficient for identification of the protein in question. However, performing a multipoint internal calibration will improve the precision of the final dataset. Furthermore, the peaks used for calibration may be missing, of poor quality or the isotope envelope may overlap with other peaks, making internal calibration next to impossible.

Solution

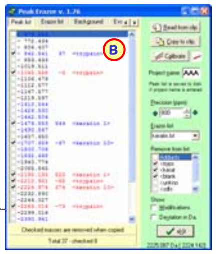
To address these problems we have created a small, user-friendly Windows-based program, **PeakErazor**, using a simple concept: By comparing all mass values from a tryptic digest against a list of known contaminants, contaminating peaks can easily be identified and removed. At the same time you can perform a multipoint calibration leading to improved precision. These features **leaves you with a much larger precision in the actual PMS search**. Furthermore, the program helps you detect new contaminants either in a given dataset or after general use.

A novel development of the program enables the calibration using the **mass defect** which results in the calibration of mass spectra that we hitherto was not able to calibrate. In general the final precision is in the 30-50 ppm region.

2 - General function of PeakErazor

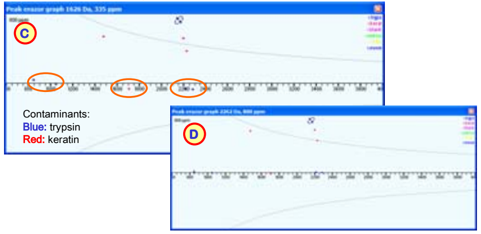


C:\Documents and Settings\Peter\My Documents\WORD\Teaching\BM131\Day4\spec2.massml (12:57 01/30/05)
Description: SSI 39.5



A. A MALDI mass spectrum is annotated, the mass list is copied to the clipboard.

B. The mass list is pasted into PeakErazor. The program compares it to the currently loaded 'Erazor list' and calculates the difference to all 'contaminants' that fall within the given precision.



C. In the separate graph window the deviations are plotted against the mass. A number of spots are seen around the x-axis, indicating a likely calibration line. The three outliers are unchecked prior to calibrating on the five remaining values.

D. After calibration the spots are within 13 ppm (which gives a norm for the database search). One of the outliers is directly on the +1 Da grey line and is thus a candidate for a wrongly assigned isotope. These outlying peaks are unchecked from the list before all unchecked mass values are copied to the clipboard for PMS.

The final protein (Calreticulin) was identified with 10 peptides (± 25 ppm). After a two-point calibration, the same protein was identified with only 9 peptides (± 30 ppm) due to a ~10 ppm offset and a slight slope relative to the multipoint calibration.

3 - Compensating for variations in the mass defect

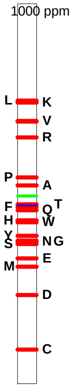
The mass defect

The mass defect is the difference between the monoisotopic and the integer mass value of a given amino acid residue, i.e. the mass defect of Alanine with a monoisotopic mass of 71.03711 Da is 0.03711 Da or 523 ppm.

Mass defect of the 20 amino acid residues in ppm:

Ala	523	Arg	648	Asn	377
Asp	234	Cys	89	Glu	330
Gln	457	Gly	376	His	430
Leu	743	Lys	741	Met	309
Phe	465	Pro	544	Ser	368
Thr	472	Trp	426	Tyr	388
Val	691				

The average mass defect is 494 ppm.



The mass defect

The distribution of the mass defect for all amino acid residues shown on a scale from 0 to 1000 ppm. The blue line indicates the average mass defect, the green line average not including Cys.

Calibrating on the mass defect

An initial attempt to calibrate on the mass defect, i.e. performing a 'straight' multipoint linear calibration on the deviation from the average mass defect, yielded a reasonable mass precision in the order of 50-80 ppm. This was better than no calibration, but prompted an investigation into the nature of the mass defect.

As the most common enzyme by far is trypsin, an enzymatic digestion of the entire Swiss-Prot database (164.000 proteins) was done *in silico*, yielding 6.55 million peptides, of which 4.1 million was in the interesting region of 500..4000 Da.

Peptides were divided into 50 Da. slots. For each slot a corrected mass defect (CMD) was calculated as

CMD = actual mass value / av. mass defect mass

Total CMD: Σ (Round (CMD) – CMD) / pepCount

The figure shows:

Green: - Peptide distribution (x 65.500)

Red: – number of peptides (in %) deviating more than 125 ppm

Blue: Total CMD x 100

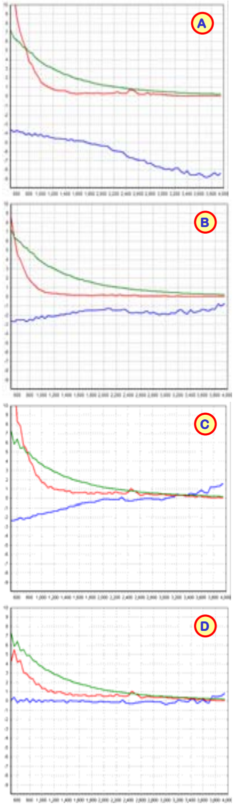


Figure A.

Deviation from mass defect calibration of tryptic peptides based on the average mass defect of 494 ppm. Less than 1% of tryptic peptides above mass 1050 Da. deviate with more than 125 ppm.

Figure B.

Omitting the mass defect of Cys from the average mass defect (448 ppm) improves the total CMD and lowers the number of peptides with large deviations.

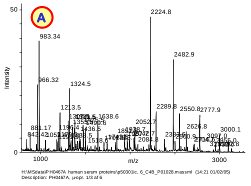
Figure C.

Adjusting the average mass defect for the database residue distribution (494 ppm) brings the total CMD to zero in the range 2000-3000 Da., but the low and high mass ranges are not improved due to the fact that every peptide has a Lys or Arg, both of which have a high mass deviation. The number of peptides with large deviations actually gets worse.

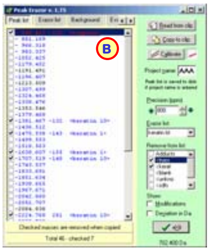
Figure D.

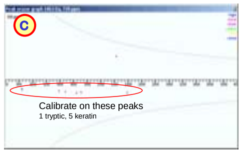
Adjusting the average mass defect factor in the low and high mass regions yields a total CMD close to zero. Also the number of peptides >125 ppm decreases considerably.

4 - Multipoint and mass defect calibration



H:\MSdata\PH0467A human serum proteins\p50301ic, 6_C4B_P01028.massml (14:21 01/02/05)
Description: PH0467A, µ-opr. 1/3 af 6



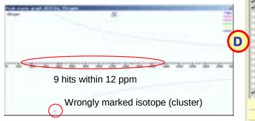


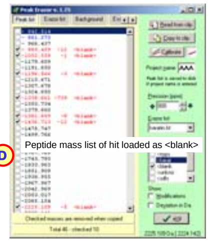
A: When the mass spectrum has been annotated (in this case using MoverZ), copy the list of monoisotopic masses to the clipboard.

B: Paste the list into PeakErazor. Note that the mass values, which fit to values in the chosen 'Erazor list' (e.g. known contaminants) will be checked and listed with deviation and id.

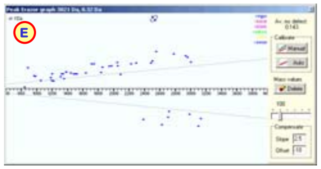
C: In the mass vs. deviation graph it is clear to see that several of the contaminants are located on a straight line.

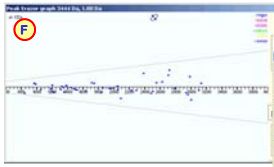
D: After deselecting the outlying peak (m/z 2224) the remainder is used for a linear calibration. After the target protein has been identified, it is digested *in silico* and the peptide mass values reloaded into PeakErazor as <blank>. The identified peptides can be seen as **9 hits within 12 ppm**. One mass value (m/z 1338) can be seen on the bottom gray line (- 1 Da) and upon inspection of the mass spectrum, it could be seen to be a wrongly marked isotope.

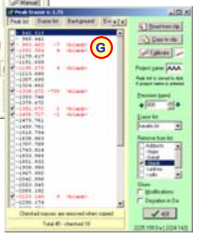



Peak Erazor v. 1.75
D
Peptide mass list of hit loaded as <blank>
Checked masses are removed when copied
Total 46 - checked 10

E: After switching the graph display to 'Mass defect view', the same raw mass list is pasted into PeakErazor. The mass defect for each peptide is now plotted relative to the mass and can be seen to form a diffuse sloping line.





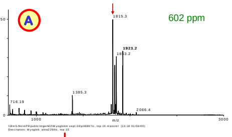


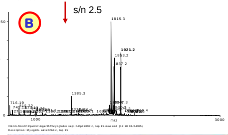
F: Performing an linear multipoint calibration on the mass defect with removal of 'outliers' (>125 ppm) yields a diffuse scattering of points centered around the x-axis. The mass list is now calibrated and can be used for PMS.

G: Reloading the target protein peptide list shows **9 hits within 10 ppm**.

5 - Isolated peptides

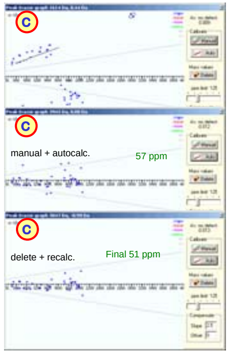
A major problem when analyzing (HPLC) isolated peptides is that your only choice for calibration is either an external calibration or calibrating on matrix peaks. However, in many cases the mass defect calibration can also be used here – it is mainly a case of finding enough (low intensity) monoisotopic mass peaks.





A. The initial mass spectrum of an HPLC separated peptide shows a peptide peak at m/z 1815. Not being micropurified, the spectrum shows a number of adduct ions. The precision is **quite poor at 602 ppm**.

B. Decreasing the signal to noise ratio to 2.5 labels a large number of peaks, most of which are peptide related, but of very low intensity. This peptide mass list is copied into PeakErazor.



C. The mass default deviation map show a distribution both above and below the x-axis. This necessitates a manual calibration, performed by selecting 'Manual' followed by a click at both ends of the low mass 'line of deviations' as shown.

D. A following 'Auto' calibration performs a linear fit to the existing data. A check of the resulting mass values shows that the precision is now 57 ppm.

E. Deleting the mass values with a mass defect >125 ppm (matrix and adduct ions) by pressing 'Delete' followed by another automatic calibration yields the final mass values which shows a **precision of 51 ppm.**

6 - Conclusion

The current program depends on a combination of the human brain to pick out visual trends in a graph followed by a linear multipoint calibration. Using the 'standard' calibration option of calibrating on contaminations a 20-50% improvement in the precision of the mass list can usually be achieved. In addition, contaminant mass values are removed from the search list, thus improving the search as the mass list gets smaller.

The program saves the mass search data as either removed or included in the search, and after several hundred spectra have been analyzed, the program enables you to identify system specific contaminants. Another functions enables you to identify common values in a smaller dataset (i.e. contaminants specific to a particular gel or when analyzing isoforms/alleles).

The most recent addition to the program, **mass defect calibration**, comes into its own in the cases where no internal calibrants or contaminants can be located in the data. By taking advantage of the fact that peptide mass values are not evenly distributed, but most values fall within ±125 ppm regions it is possible to calibrate any peptide containing mass spectrum given that a sufficient number of peptides are present. As a side effect, most non-peptide mass values can be removed from the list, as they will have a mass defect larger than 125 ppm.

The mass defect calibration does not work if the spectrum is heavily contaminated with non-peptide peaks (i.e. matrix, detergent or adduct ions). These cases are usually immediately recognized, as the spread of the mass defect gets larger than 125 ppm and it is not possible to perform a 'stable' calibration. Few non-peptide mass values are easily removed and recalibration is just the press of a button. Furthermore, the higher the number of peptides, the more accurate the calibration. In praxis you need ~15 mass values for a proper calibration. In a few cases (in our experience <5%) the method fails and we have to resort to external calibration.

Reloading the theoretical peptide mass list of the target protein enables you to identify wrongly assigned peptides (larger than average deviation) and helps to locate potentially modified peptides.

8 - Where to find PeakErazor

The PeakErazor program can be downloaded for free from

http://welcome.to/gpmaw

Please go to the 'download' section of the web site.

If you have any problems or have suggestions for improvements, please contact the author on php@bmb.sdu.dk.