High Probability MS Identification of Phosphopeptides by their Mass Defect

Erol E. Gulcicek[2], Mark Shifman[1], Perry Miller[1], Christopher M. Colangelo[2], Kathryn L. Stone[2], Matthew Berberich[2], TuKiet T. Lam[2], Can Bruce[1]

Yale University

Center for Medical Informatics[1], W.M. Keck Foundation Biotechnology Resource Laboratory[2]

Overview / Introduction

The mass defect (the difference between the actual and nominal monoisotopic mass) of the phosphate group is smaller than that of most atoms present in proteins. When the mass defects of tryptic peptides derived from the human proteome are plotted against their masses, phosphopeptides tend to fall off the regression line. In this study, we explored whether this characteristic of phosphorylated peptides can be used to distinguish them with high probability from unphosphorylated peptides.

Background

Various techniques have been developed in recent years that uses the mass defect principle to facilitate the identification and quantitation of peptides (see (1) for a review). We demonstrate the concept of fractional mass to be able to detect phosphopeptides apart from their more abundant unphosphorylated counterparts in a mass spectrometer. When using a tandem MS-MS instrument for biological mass spectrometry, it is difficult to determine if a particular peptide is phosphorylated in the "survey mode" of operation. As a result, valuable time is spent using the mass spectrometer for sequence determination of unphosphorylated peptides that are generally the most abundant peptides. This time commitment may be in excess of the time window in which a transient species may be available for analysis, particularly if the MS-MS is coupled to an online chemical separation technique such as liquid chromatography or capillary electrophoresis.

The concept uses theoretical negative mass shift behavior of phosphopeptides and assigns probabilities to these peptides (Bruce et al. 2006) that can be distinctively selected for MS/MS processing before the more abundant unphosphorylated peptides. This idea can be implemented in real time to first rank order probabilities of peptides for being a phosphopeptide and preferentially decide the order of acquisition for MS/MS from a list of possible peptides. The assigned probability would also increase the confidence and the probability of identification of many phosphopeptides.

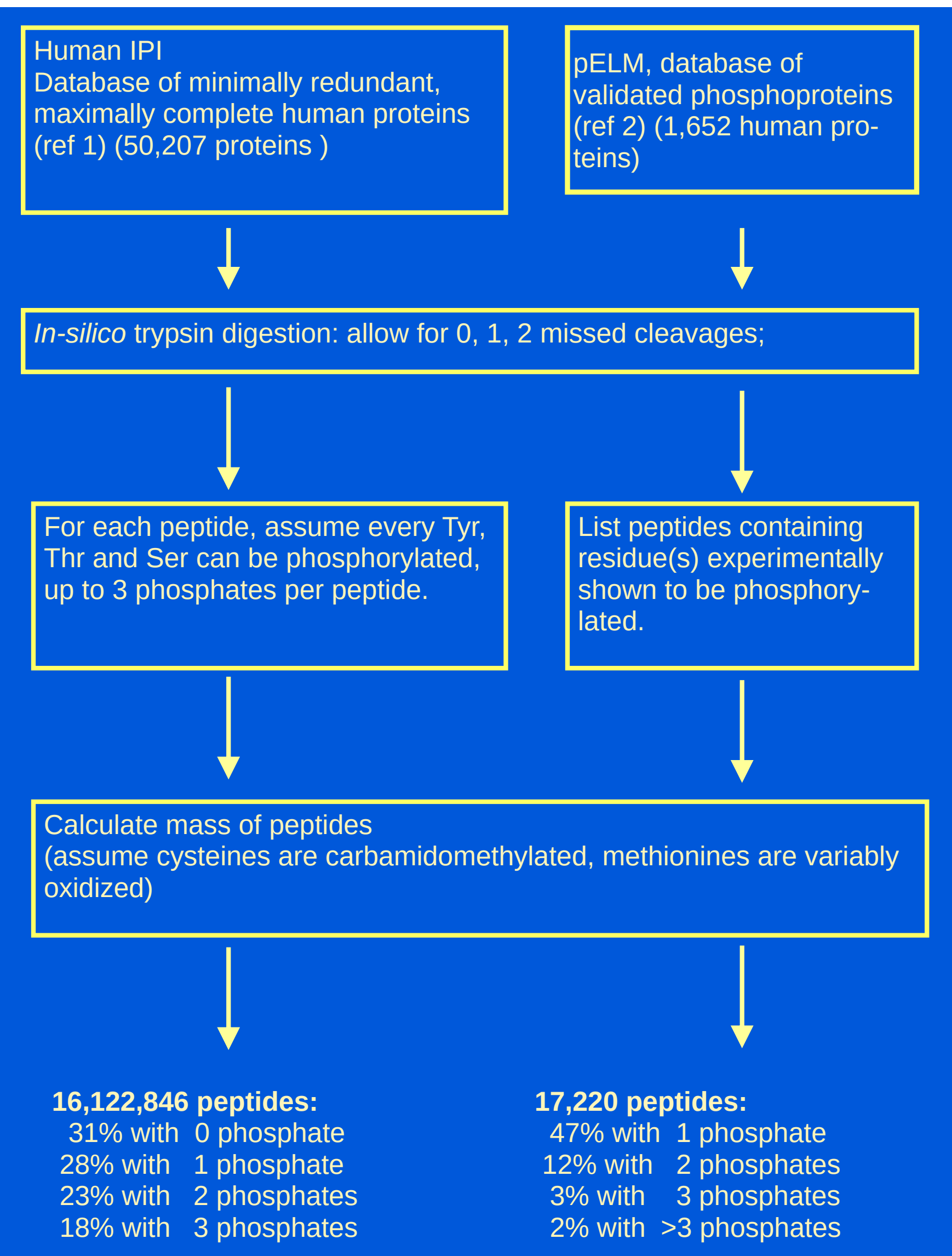
	Mass	Mass Defect
^{12}C	12.0000000	+0.0000000
^{14}N	14.0030740	+0.0030740
^{1}H	1.0078250	+0.0078250
^{16}O	15.9949146	-0.0050854
^{32}S	31.9720707	-0.0279293
^{31}P	30.9737615	-0.0262385
HPO_3	79.9663304	-0.0336696
Ser	87.0320284	+0.0320284
SerSer	175.0713333	+0.0713333
pSer	168.0056353	+0.0056353

Mass Defect: difference between the actual and nominal monoisotopic mass.

Fractional Mass: Mass value to the right of the decimal point.

Up to 0.999…, Fractional Mass and Mass Defect are equal. Above ~1500 Daltons, mass defect is >1.0, while fractional mass cycles back to stay in the range of [0 - 0.99...].

Methods



Human IPI
Database of minimally redundant, maximally complete human proteins (ref 1) (50,207 proteins)

pELM, database of validated phosphoproteins (ref 2) (1,652 human proteins)

In-silico trypsin digestion: allow for 0, 1, 2 missed cleavages;

For each peptide, assume every Tyr, Thr and Ser can be phosphorylated, up to 3 phosphates per peptide.

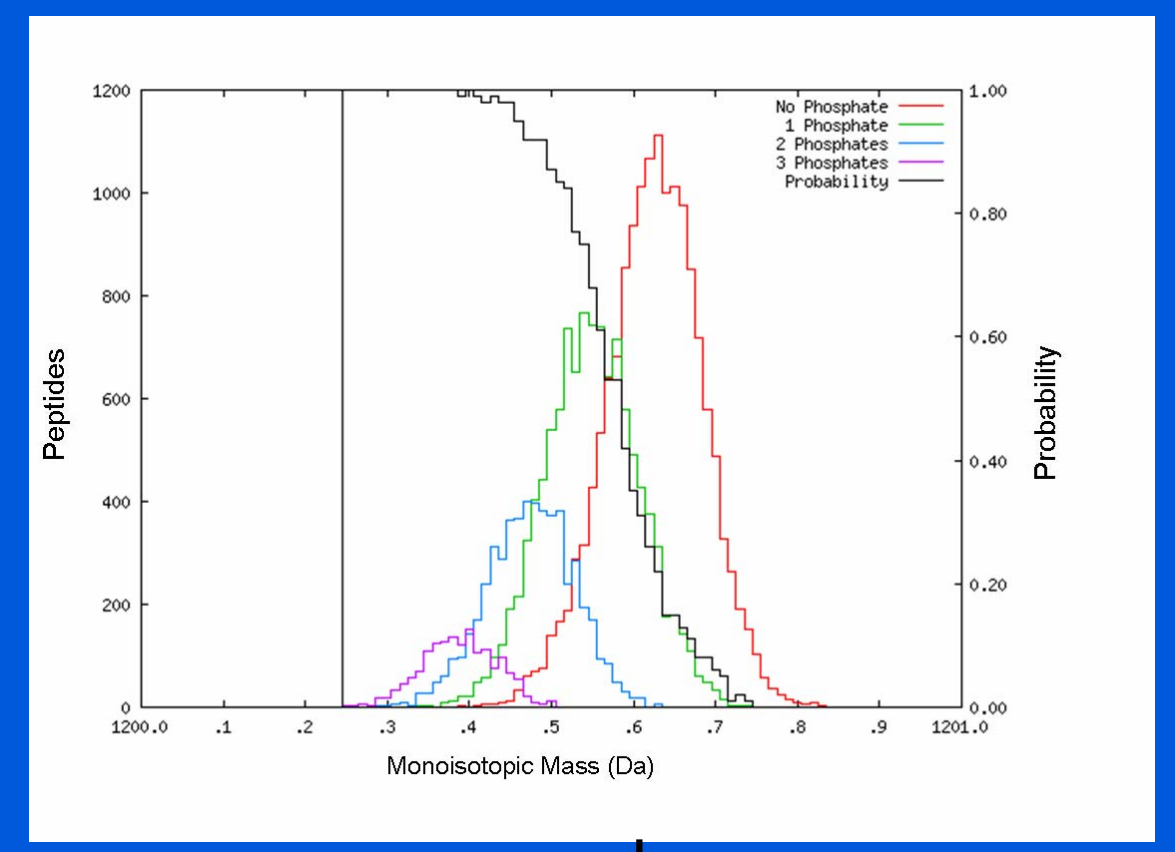
List peptides containing residue(s) experimentally shown to be phosphorylated.

Calculate mass of peptides
(assume cysteines are carbamidomethylated, methionines are variably oxidized)

16,122,846 peptides:
31% with 0 phosphate
28% with 1 phosphate
23% with 2 phosphates
18% with 3 phosphates

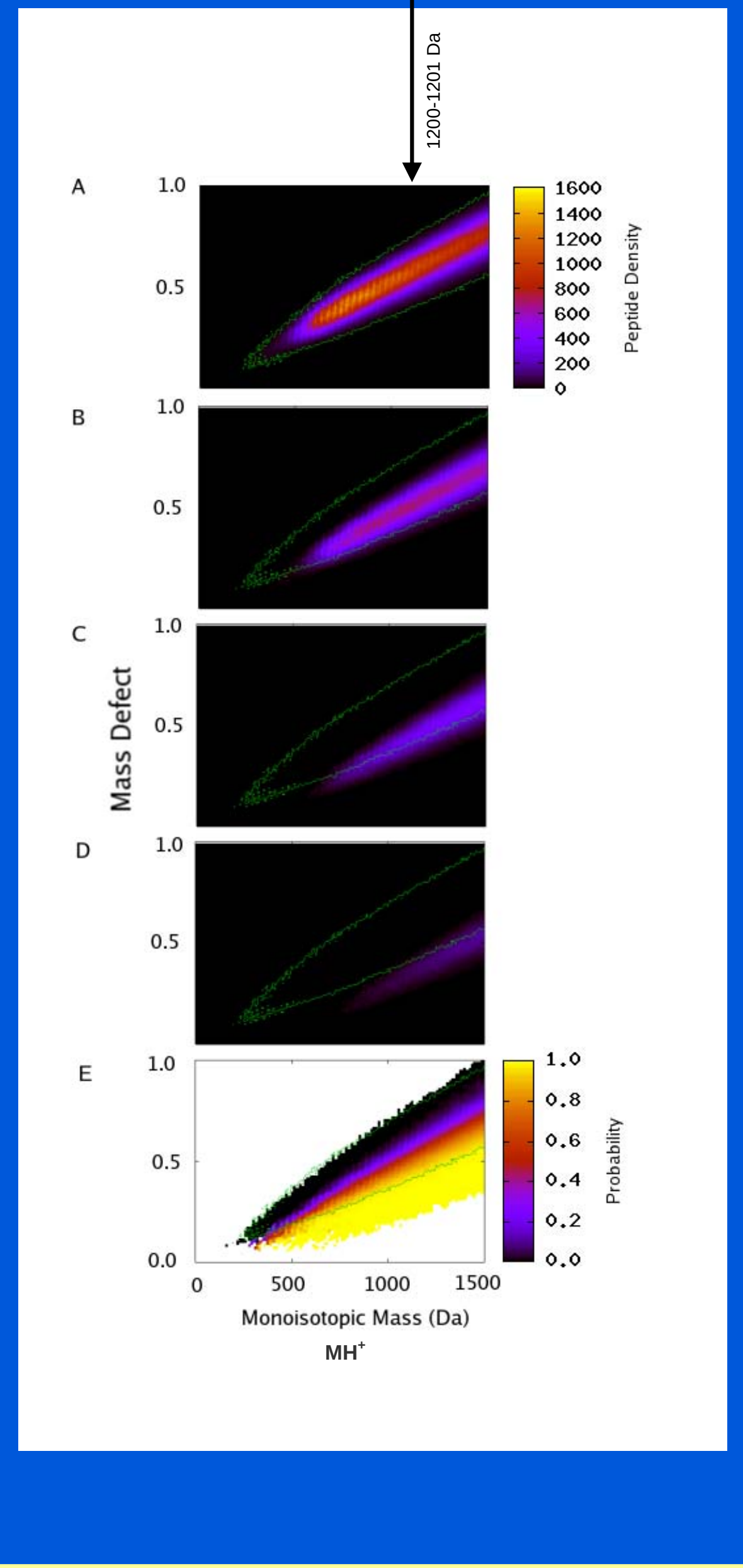
17,220 peptides:
47% with 1 phosphate
12% with 2 phosphates
3% with 3 phosphates
2% with >3 phosphates

Results



Tryptic peptides with the same nominal mass cluster around narrow mass ranges. Depending on the number of phosphates, the tryptic peptide distribution shifts allowing the calculation of probability of a phosphopeptide as a function of fractional mass.

Figure 1. Distribution of number of peptides and phosphorylation probability for tryptic peptides with 0, 1, 2, and 3 phosphates, shown here for the mass range 1200 to 1201. Each bin size is 0.01 Da.



Density Plots for phosphopeptides shift down with increasing number of phosphate groups on Mass Defect vs. Mass plot.

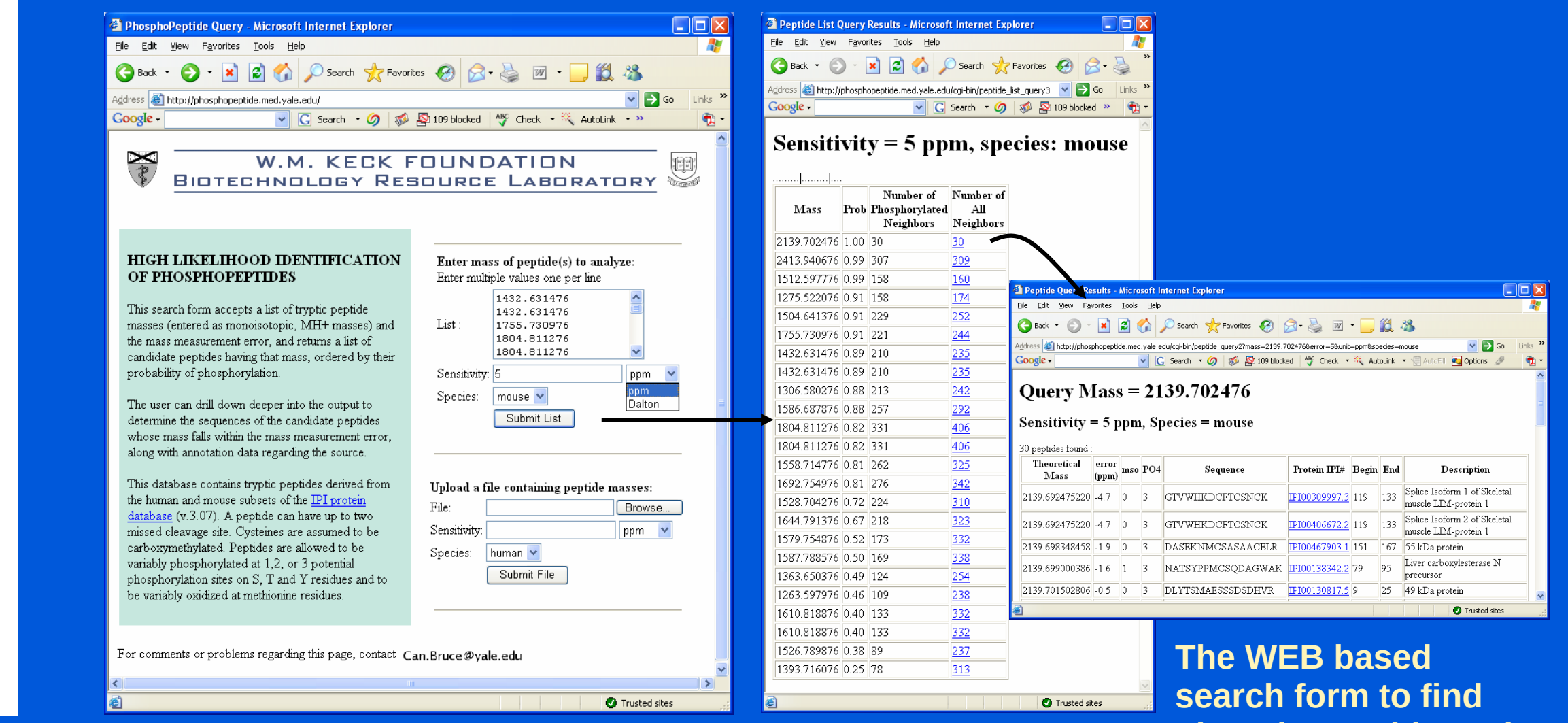
Figure 2. Density map of peptide numbers and phosphorylation probability for tryptic peptides with 0, 1, 2 and 3 phosphates. **A-D,** Density map of peptides with 0, 1, 2 and 3 phosphates respectively, as a function of Mass Defect and Mass. Colors denote number of peptide within one pixel of the map, each pixel having dimension of 10 Da Mass x 0.01 Da Mass Defect. Color map in panel A applies to panels B-D as well.

Green contour in Panel A shows boundary or region having ≥ 10 peptides per pixel and is reproduced in panels B-E for reference.

E, Phosphorylation probability distribution as a function of Mass Defect and Mass.

Probability of phosphorylation = (N1+N2+N3) / (N0+N1+N2+N3)

where N0, N1, N2 and N3 are the number of peptides with 0, 1, 2, and 3 phosphates, respectively, within each pixel.



PhosphoPeptide Query - Microsoft Internet Explorer
Address http://phosphopeptide.med.yale.edu/

W.M. KECK FOUNDATION
BIOTECHNOLOGY RESOURCE LABORATORY

HIGH LIKELIHOOD IDENTIFICATION OF PHOSPHOPEPTIDES

This search form accepts a list of tryptic peptide masses (entered as monoisotopic, MH+ masses) and the mass measurement error, and returns a list of candidate peptides having that mass, ordered by their probability of phosphorylation.

The user can drill down deeper into the output to determine the sequences of the candidate peptides whose mass falls within the mass measurement error, along with annotation data regarding the source.

This database contains tryptic peptides derived from the human and mouse subsets of the IPI protein database (v.3.07). A peptide can have up to two missed cleavage site. Cysteines are assumed to be carboxymethylated. Peptides are allowed to be variably phosphorylated at 1,2, or 3 potential phosphorylation sites on S, T and Y residues and to be variably oxidized at methionine residues.

Enter mass of peptide(s) to analyze:
Enter multiple values one per line

List: 1432.631476
1432.631476
1755.730976
1804.811276
1804.811276

Sensitivity: 5 ppm
Species: mouse

Submit List

Upload a file containing peptide masses:
File:
Sensitivity: ppm
Species: human

Submit File

For comments or problems regarding this page, contact Can.Bruce@yale.edu

Peptide List Query Results - Microsoft Internet Explorer
Address http://phosphopeptide.med.yale.edu/cgi-bin/peptide_list_query3

Sensitivity = 5 ppm, species: mouse

Mass	Prob	Number of Phosphorylated Neighbors	Number of All Neighbors
2139.702476	1.00	30	30
2413.940676	0.99	307	309
1512.597776	0.99	158	160
1275.522076	0.91	158	174
1504.641376	0.91	229	252
1755.730976	0.91	221	244
1432.631476	0.89	210	235
1432.631476	0.89	210	235
1306.580276	0.88	213	242
1586.687876	0.88	257	292
1804.811276	0.82	331	406
1804.811276	0.82	331	406
1558.714776	0.81	262	325
1692.754976	0.81	276	342
1528.704276	0.72	224	310
1644.791376	0.67	218	323
1579.754876	0.52	173	332
1587.788576	0.50	169	338
1363.650376	0.49	124	254
1263.597976	0.46	109	238
1610.818876	0.40	133	332
1610.818876	0.40	133	332
1526.789876	0.38	89	237
1393.716076	0.25	78	313

Peptide Query Results - Microsoft Internet Explorer
Address http://phosphopeptide.med.yale.edu/cgi-bin/peptide_query2?mass=2139.702476&error=5&unit=ppm&species=mouse

Query Mass = 2139.702476

Sensitivity = 5 ppm, Species = mouse

30 peptides found :

Theoretical Mass	error (ppm)	mso	PO4	Sequence	Protein IPI#	Begin	End	Description
2139.692475220	-4.7	0	3	GTVWHKDCFTCSNCK	IPI00309997.3	119	133	Splice Isoform 1 of Skeletal muscle LIM-protein 1
2139.692475220	-4.7	0	3	GTVWHKDCFTCSNCK	IPI00406672.2	119	133	Splice Isoform 2 of Skeletal muscle LIM-protein 1
2139.698348458	-1.9	0	3	DASEKNMCSASAACELR	IPI00467903.1	151	167	55 kDa protein
2139.699000386	-1.6	1	3	NATSYPPMCSQDAGWAK	IPI00138342.2	79	95	Liver carboxylesterase N precursor
2139.701502806	-0.5	0	3	DLYTSMAESSSDSDHVR	IPI00130817.5	9	25	49 kDa protein

http://phosphopeptide.med.yale.edu/

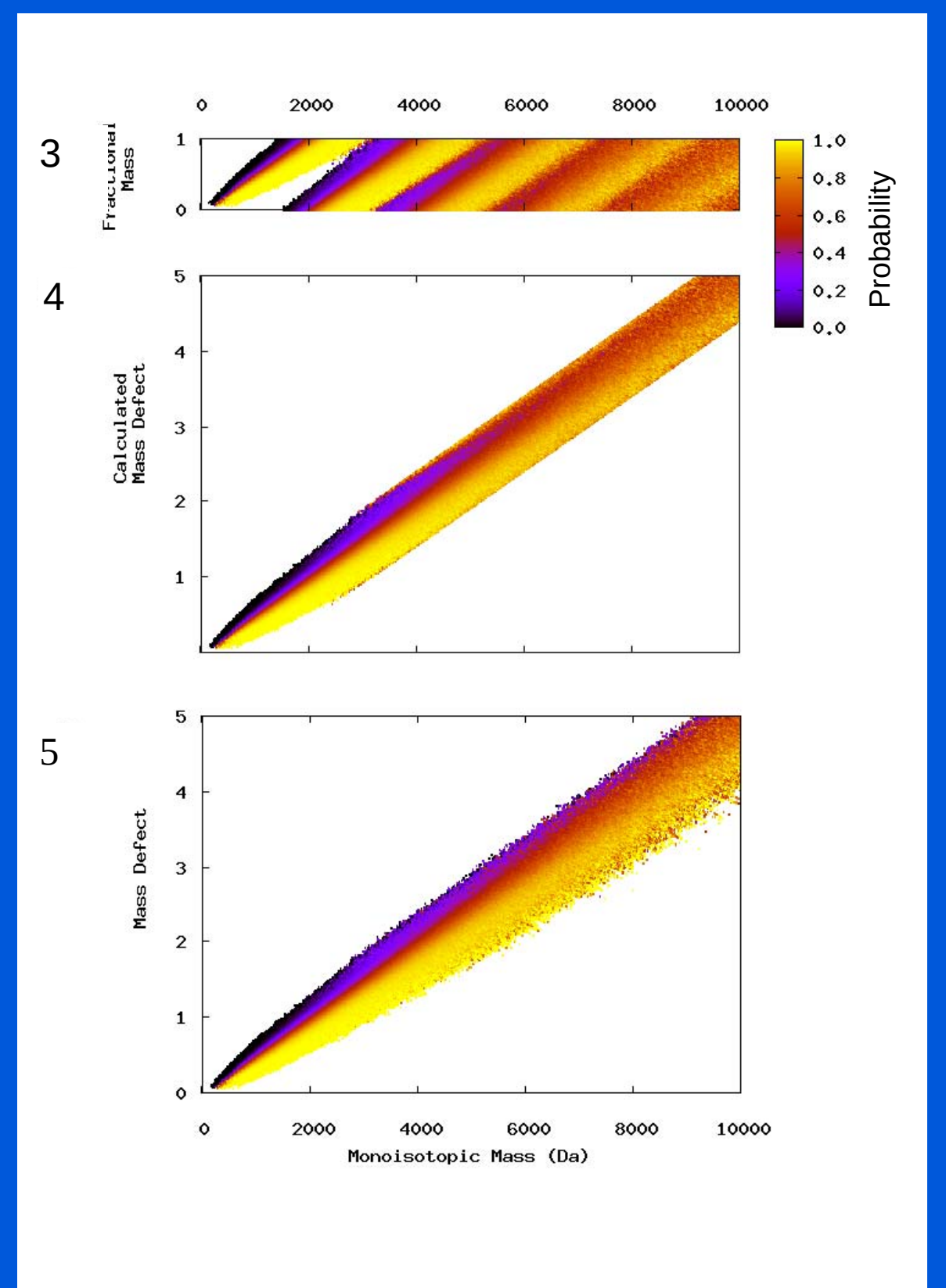
The WEB based search form to find phosphopeptide probabilities

We developed a transformation function to calculate the mass defect of a peptide from its mass:

Calculated Mass Defect = int(aM+b - int(aM+b) - F) + int(aM+b) + aM+b + F

where the function int(x) returns the greatest integer smaller than x, M is monoisotopic mass, F for fractional mass (that is, M-int(M)) and a and b are constants.

Using this formula, 98.8% of all peptides have the correct calculated Mass Defect value, The mismatches occurring at the edges of the data cloud where the addition of the correct integer value to the fractional mass is more prone to error (see Fig 5 and 6). These errors primarily occur above 4000 Da, while most real mass spectrometric peaks are observed below this mass.

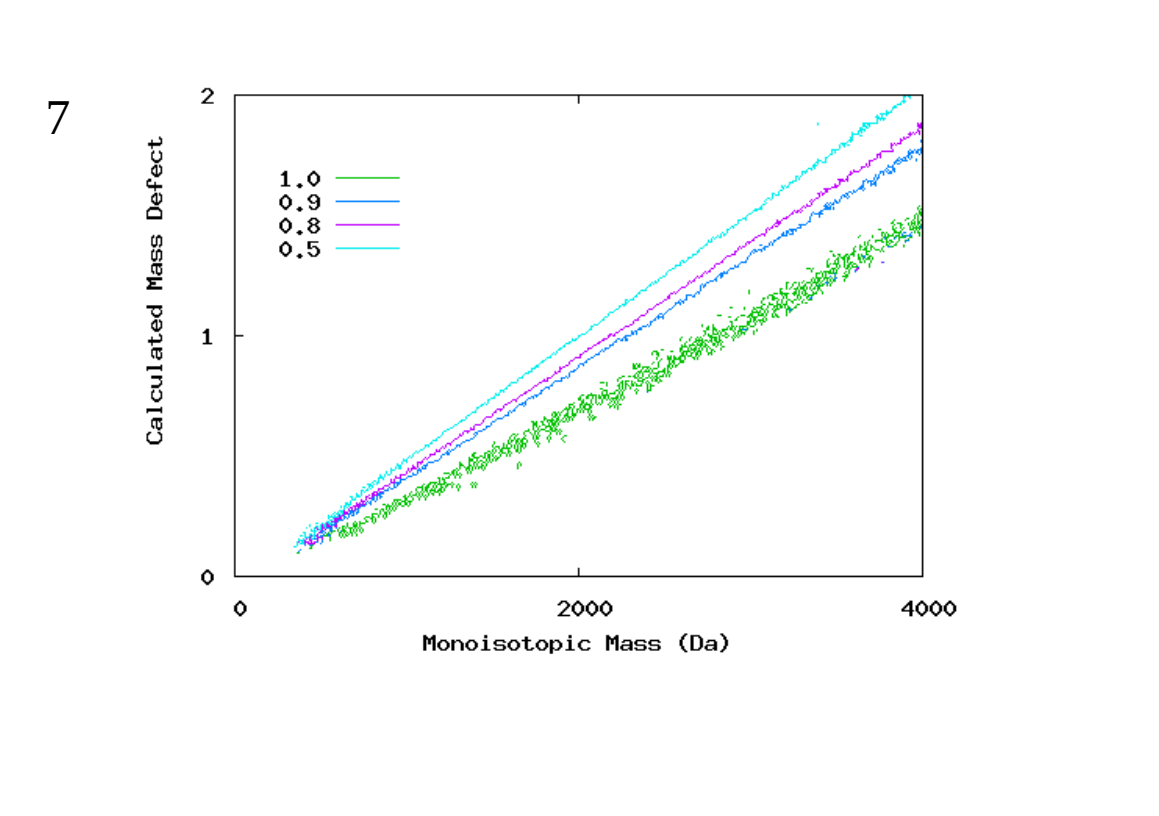


Phosphorylation probability distribution maps.

Figure 3. Phosphorylation probability distribution map as a function of Monoisotopic Mass and Fractional Mass. Color Map shows probability of phosphorylation for each pixel of 10 Da Mass x 0.01 Da Fractional Mass.

Figure 4. Phosphorylation probability distribution map as a function of Monoisotopic Mass and calculated Mass Defect.

Figure 5. Phosphorylation probability distribution map as a function of Monoisotopic Mass and true Mass Defect.

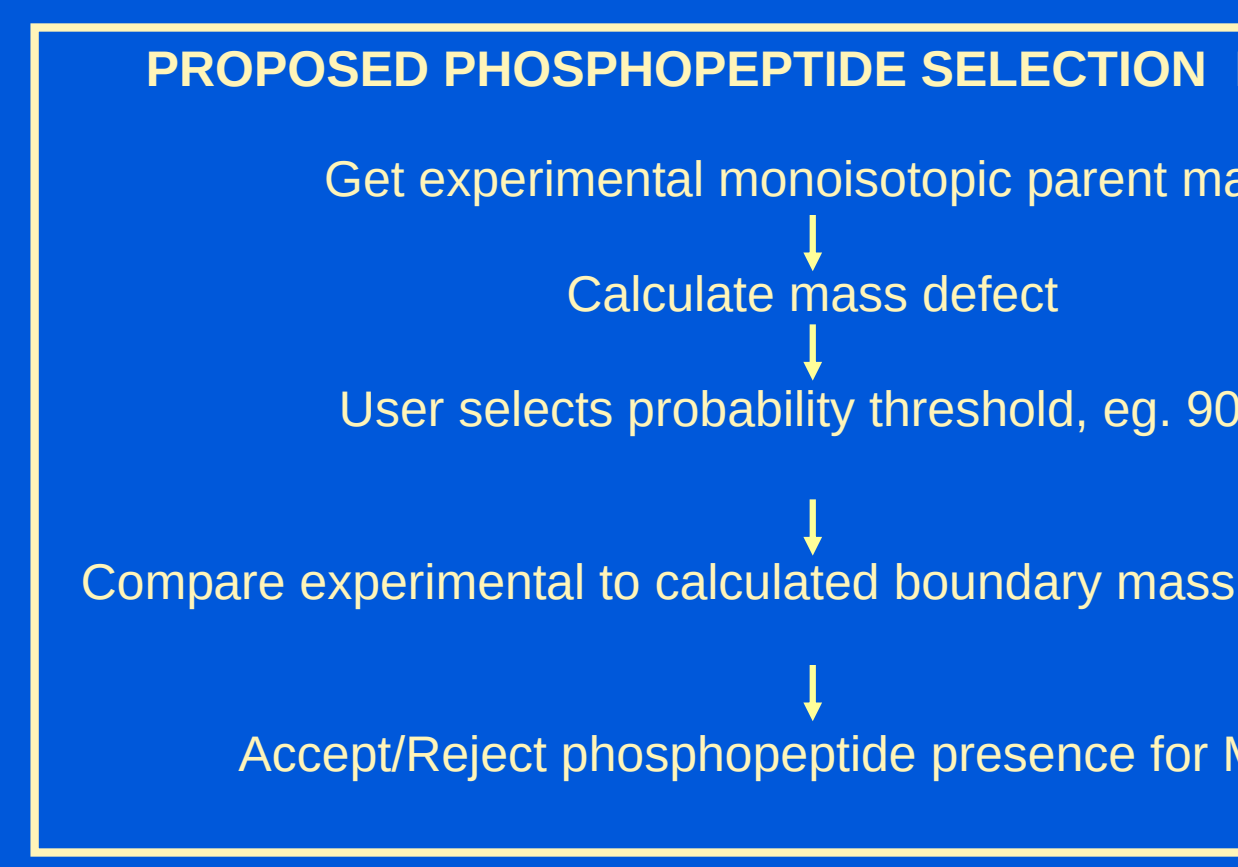


Phosphorylation probability contour maps

Figure 7. Probability distribution data in Fig. 5, above, re-plotted as a contour map to show linearity of selected contour lines. Note different scale of the graph axes.

Each of these contour lines can be fitted by regression lines (y=ax+b) . For example, the parameters for the p=0.9 line are
a = 0.000457± 0.0000006
b = -0.0448 ± 0.0029

Peptides that fall below the blue line in the figure have a probability >0.9 of having 1, 2, or 3 phosphate groups.



PROPOSED PHOSPHOPEPTIDE SELECTION PROCESS

Get experimental monoisotopic parent mass

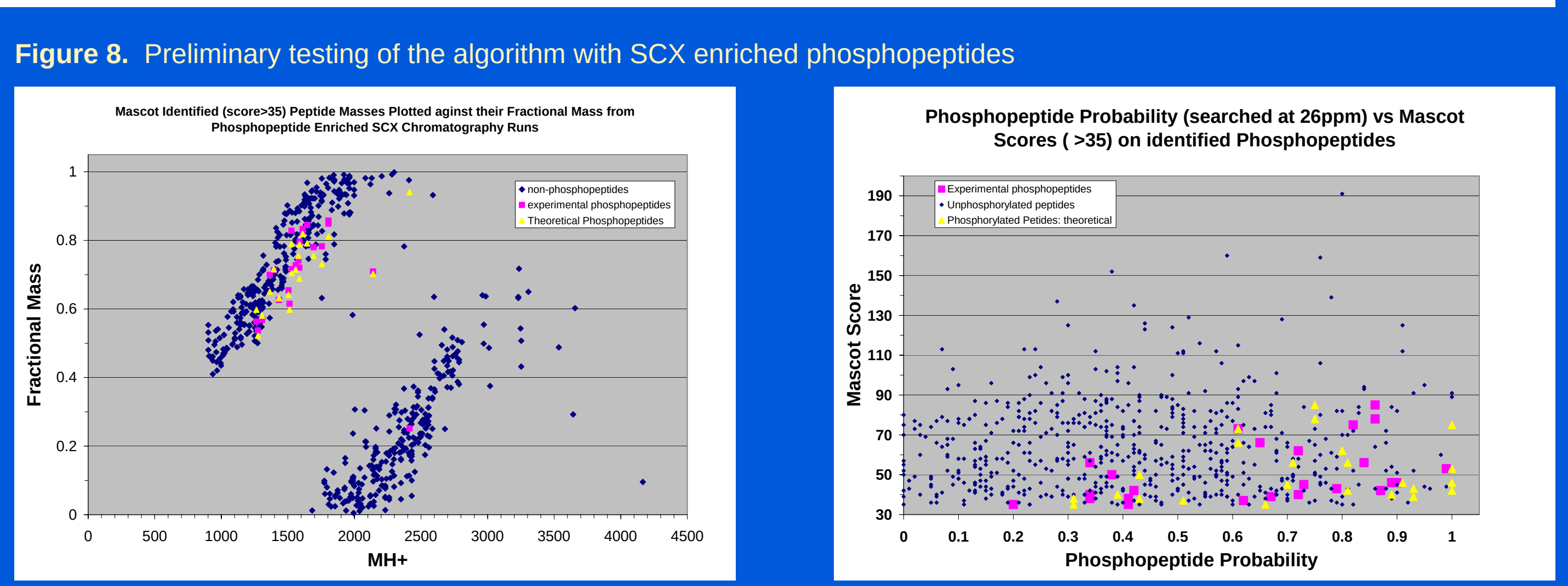
Calculate mass defect

User selects probability threshold, eg. 90%

Compare experimental to calculated boundary mass defect values

Accept/Reject phosphopeptide presence for MS/MS

Figure 8. Preliminary testing of the algorithm with SCX enriched phosphopeptides

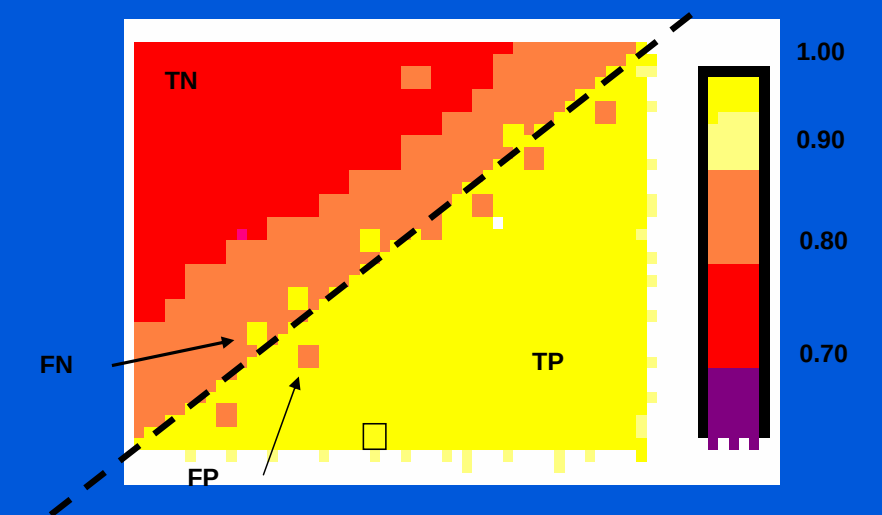


Digested Protein Sample from mouse kidney cell lines (225μg) is enriched for phosphopeptides using SCX chromatography. 600 peptides were identified from the 24 early fractions with a MASCOT score of 35 and higher. The early SCX fractions were combined to run in 4 RP LC MS/MS runs with our QSTAR mass spectrometer. Of the 600 identified peptides 24 were phosphorylated. The above two figures show the fractional mass (left) and the phosphopeptide probability distributions (right) for the 600 identified peptides. The complete experiment with 14 LC-MS/MS runs resulted in 888 protein ID's with ProteinProphet Probability of 0.9 and higher. Of these, 37 were phosphoproteins.

The p=0.9 line is highly reliable in identifying phosphopeptides up to a mass measurement accuracy of 20 ppm.

Table 2. Positive Predictive Value, TP/(TP+FP), of p=0.9 Line as a Function of Simulated Mass Measurement Error

Error	1 Phos	2 Phos	3 phos	1,2 or 3 Phos
0 ppm	80.6%	89.1%	89.5%	95.4%
1 ppm	80.6%	89.1%	89.5%	95.4%
5 ppm	80.4%	88.8%	89.2%	95.3%
20 ppm	76.9%	85.1%	85.0%	93.6%
100 ppm	55.2%	53.5%	46.0%	76.4%



Each peptide's mass was modified by an error term 50 times and each time its calculated mass defect was compared with the L $_{0.9}$ line. Numbers (except for the 0 ppm error) are the averages of these 50 independent determinations. Standard deviations are less than 0.1% of the means.

TP (true positive), phosphopeptide and below the line; FP (false positive), unphosphorylated peptide and below the line

Conclusions and Future Work

- The mass defect of tryptic peptides can be calculated with high accuracy based on their mass.
- Peptides can be classified with a simple formula for their phosphorylation probability, based on their mass and their calculated mass defect.
- ~41% of potential phosphopeptides (with 1,2 or 3 phosphates) can be identified with a probability >0.90 for peptides of Mass < 4000 Da.
- The categorization of a peptide as being phosphorylated has a positive predictive value >0.96 for Mass < 4000 Da.
- The categorization of a peptide as being phosphorylated is relatively insensitive to mass measurement accuracy up to 20 ppm.
- Experimentally validated phosphopeptides are similarly partitioned by the p=0.9 line as all potential phosphopeptides derived from the human proteome.
- An algorithm based on this formula can be used to automatically route such peptides to MS/MS analysis. *(Yale University Office of Cooperative Research Invention tracking # 4048)*

References

(1) Kersey, et al. (2004) *Proteomics 4*, 1985-1988.
(2) Diella, et al. (2004) *BMC Bioinformatics 5*, 79.
(3) Schneider and Hall (2005) Drug Discovery Today, 10:353-363.
(4) Bruce, et al. (2006) Anal Chem. (Accepted; Published on Web 05/13/06)

Acknowledgements

This project has been funded in part with federal funds from NHLBI contract (N01-HV-28186) and NIDA Neuroproteomics grant 1 P30 DA018343. CB is funded by NIH grant T15 LM07056 from the National Library of Medicine.