Comprehensive Analysis of a Published Dataset: Identifying more of your data using the latest Protein Prospector

Peter R. Baker, Robert J. Chalkley , Katalin F. Medzihradszky and Alma L. Burlingame

Mass Spectrometry Facility, Dept. of Pharmaceutical Chemistry, University of California, San Francisco, USA

Summary

A public dataset (Klimek et al 2007):

- File: QS20060131_S_18mix_02.wiff
- Mixture of roughly 30 proteins.
- Tryptic digest (microwave-assisted)
- 1867 spectra

The aim was to use the tools in the latest version (v5.0) of Protein Prospector (freely available on the web at http://prospector 2.ucsf.edu) to try to identify as many as possible of the spectra.

Experimental

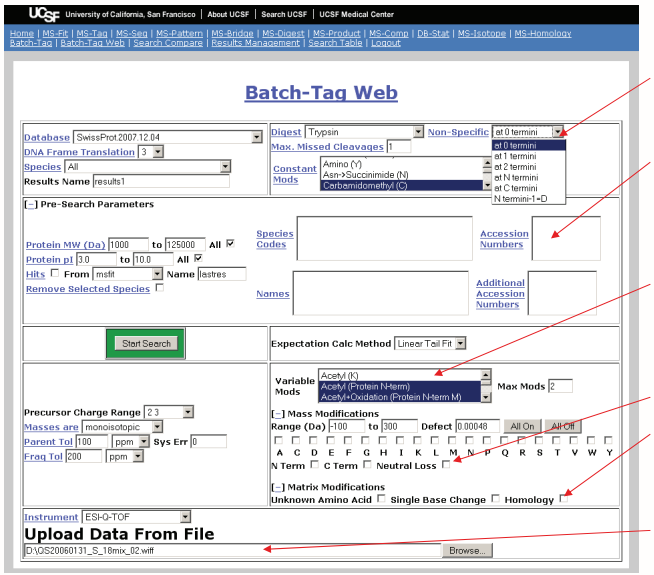
Database searching is performed as a 2 step process.

1) First the proteins present are identified using very strict criteria.
2) Data are searched against the list of identified proteins using:
   - A selection of common variable modifications.
   - Mass modification searching (Baker et al 2006)
   - No enzyme specificity.

Semi-manual analysis was then performed using various options and programs within Protein Prospector:

- The precursor data was inspected for each spectrum to determine the correct precursor m/z and charge.
- Each spectrum was also individually searched using MS-Tag from File to allow checking for multiple hits and adjustment of the precursor m/z and charge.
- A separate search was done to look for up to 2 amino acid substitutions using the homology tick box. Any substitutions identified were checked by searching NCBInr and UniprotKB using MS-Pattern.

Batch-Tag Options in PPv5



Non-specific options (No enzyme also available on Digest menu)

Search can be limited by accession number

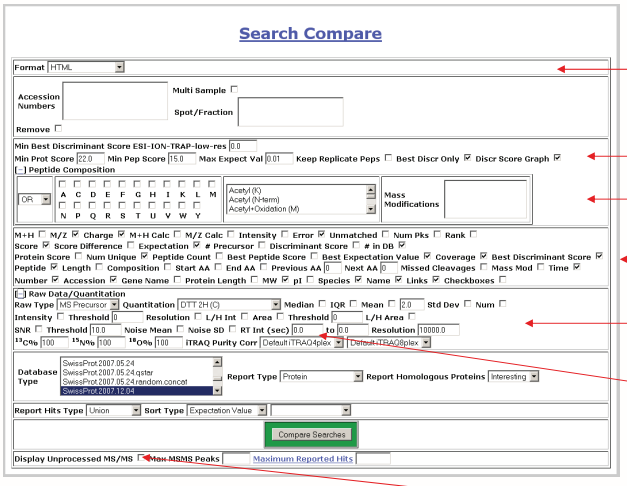
Mass modification options can be combined with variable modifications

Neutral loss option

Homology option

Wiff file can be uploaded and will be automatically centroided. The raw data precursor and MSMS data can thus be viewed.

Search Compare Options in PPv5



Output format can be HTML or tab delimited text

Score filters

Composition filters

Report Columns

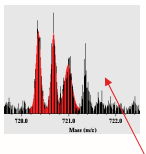
Raw data/quantitation options

Option that allows averaging of the precursor data

If this option is selected all subsequent MSMS spectra will be displayed using all the peaks in the centroid file

Links from Search Compare Peptide Report

720.3563^{+3}

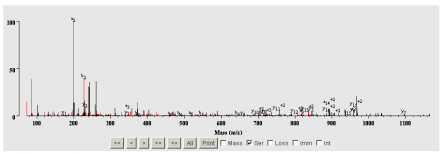


Precursor scan can be accessed here

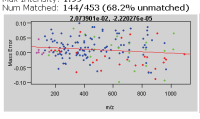
m/z	z	Da	Peptide	S	Start	End	Score	Expect	# in DB
488.2908	3	0.021	M(Met-loss)TM(Oxidation)ITDSLAVVLQR	69.18	1	14	20.5	0.40	1
723.9222	2	0.019	M(Met-loss)TMITDSLAVVLQR	70.57	1	14	54.8	1.1e-5	1
629.9550	3	0.021	RDWENPGVTQLNR	93.8	15	27	35.6	4.6e-4	1
667.8464	2	0.021	RDWENPGVTQL	45.01	15	25	37.0	0.052	1
714.8691	2	0.022	DWENPGVTQLNR	99.96	16	27	40.4	4.0e-4	1
579.7914	2	0.016	DWENPGVTQL	52.35	16	25	39.3	0.0045	1
418.2364	3	0.012	LAAHPPFASWR	42.9	28	38	23.0	0.081	1
680.8521	2	0.020	FAWFPAPEAVPE	73.3	61	72	34.4	0.0017	1
742.3813	2	0.028	ADTVVPSNWQMH	49.26	82	94	25.9	0.047	1
673.8522	2	0.028	ADTVVPSNWQM	59.6	82	93	34.1	0.0037	1
675.7775	2	0.016	HGYDAPIYTN	28.84	94	103	34.4	0.029	1
1005.5287	2	0.047	SLTFNVDESWLQEGQTR	54.33	125	141	46.5	4.6e-6	1
905.4540	2	0.030	TFNVDESWLQEGQTR	54.83	127	141	54.5	1.2e-6	1
791.3776	2	0.012	NVDESWLQEGQTR	40.76	129	141	52.8	5.3e-6	1
724.3775	2	0.033	VDESWLQEGQTR	39.17	130	141	36.0	0.0010	1
669.6743	3	0.011	IIFDGVNSAFHLWC(58.0278)NGR	71.71	142	158	45.4	2.4e-6	1
866.8770	2	0.026	IIFDGVNSAFHL	70.66	142	153	34.0	0.021	1
759.9184	2	0.028	IIFDGVNSAFHLW	77.12	142	154	20.2	0.29	1
669.3632	3	0.035	IIFDGVNSAFHLWC(Carbamidomethyl)NGR	71.18	142	158	33.0	0.0038	1
720.3563	3	0.0016	IIFDGVNSAFHLWC(210.1008)NGR	73.5	142	158	23.4	0.17	1
720.0217	3	5.0e-4	IIFDGVNSAFHLWC(209.1003)NGR	72.24	142	158	45.5	3.5e-5	1
634.2733	3	0.024	WVGYGQDSR	26.42	159	167	32.3	0.0031	1

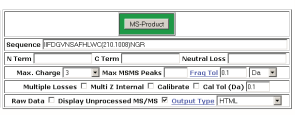
Search Compare report sorted by start amino acid shows several cysteine modifications

IIFDGVNSAFHLWC(210.1008)NGR^{+3}

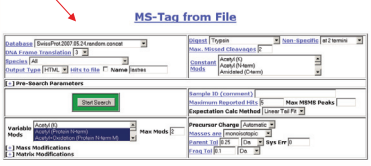


Max Intensity: 1.99
Num Matched: 144/453 (68.2% unmatched)



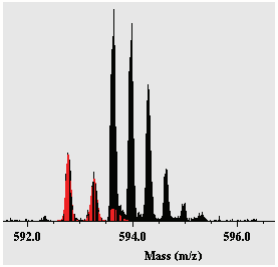


Spectra can be searched individually with different parameters



The peptide match can be viewed in MS-Product. Users have control of various aspects of how the match is displayed

Contributions from Neighboring Peaks

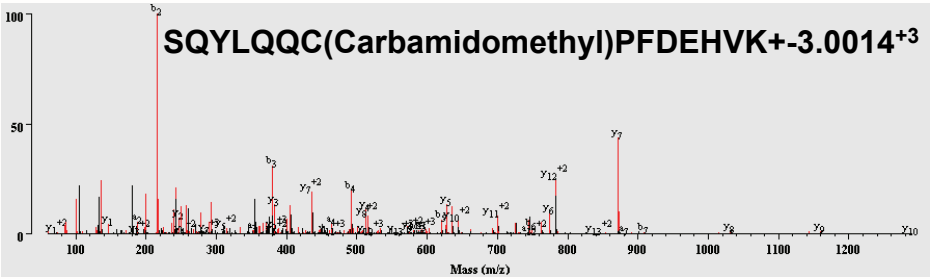


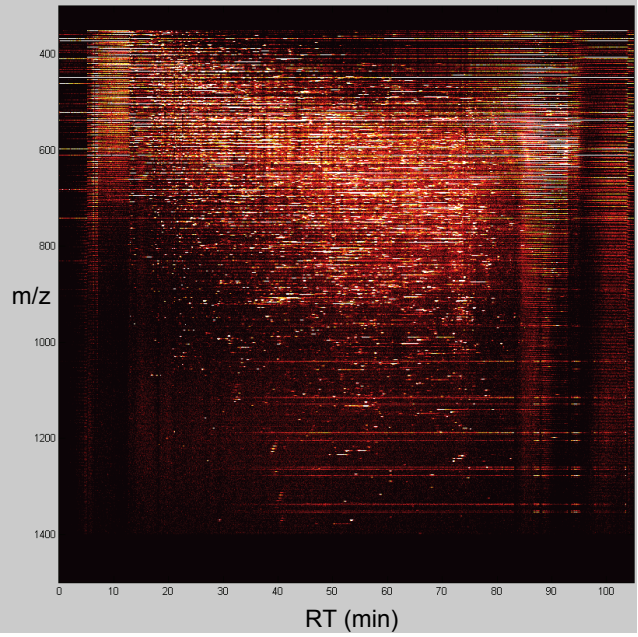
• A lot of the MSMS spectra contain more than one component. These are either from a closely co-eluting peptide or chemical background peaks.

•60 spectra had two identifiable peptides.

• The centroiding script believes the precursor is doubly charged peak (in red on the left), but most of the fragmentation is from the +3 peak.

• Prospector will find the +3 peak via the neutral loss option in mass modification searching. Around 140 similar extra hits were found via the neutral loss loss option representing cases where the major fragmentation is not at the m/z recorded in the centroid file.





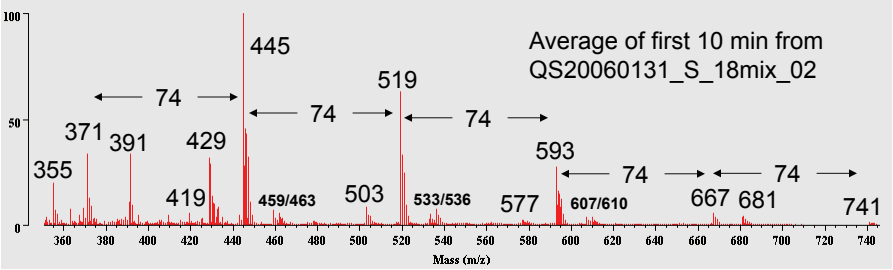
Background Ions

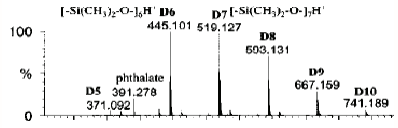
A 3D plot of the raw data clearly shows that there are no peptides in the first 12 min and the last 19 min.

Background ions can be seen as horizontal lines on this plot, the most intense being at m/z 445.

Protein Prospector can produce these plots in collaboration with MATLAB.

Background Ions



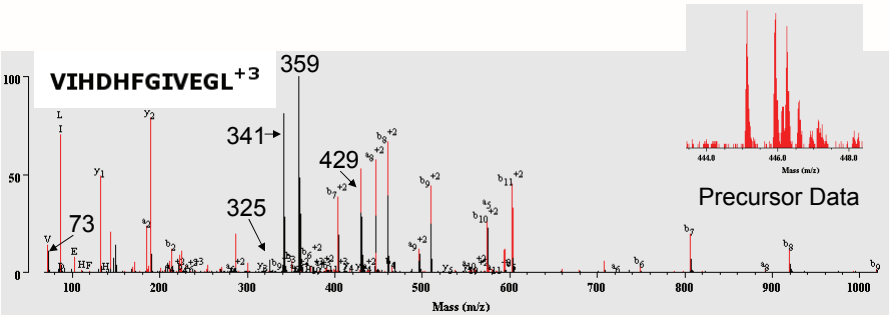


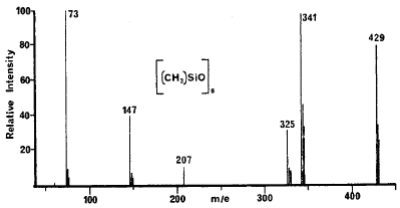
ESI spectrum showing background peaks (Schlosser and Engert 2003)

355, 371 – D5 decamethylcyclopentasiloxane
429, 445, 463 – D6 dodecamethylcyclohexasiloxane
503, 519, 536 – D7 tetradecamethylcycloheptasiloxane
577, 593, 610 – D8 hexadecamethylcyclooctasiloxane
651, 667 – D9 octadecamethylcyclononasiloxane
725, 741 – D10 eicosamethylcyclodecasiloxane
Lower peak is loss of methane (-16 Da)
Higher peak is an ammonium adduct (+17 Da)

391 – Dioctyl pthlalate

MSMS Spectrum Contaminated with D6





EI spectrum of D6 (Pickering et al 1975)

There was some background component in at least 84 spectra.

Is it worth filtering out these masses from the peaklist?

Non-Specific and Missed Cleavages

• Prospector now has a comprehensive range of options for searching for enzyme non-specific cleavage peptides.

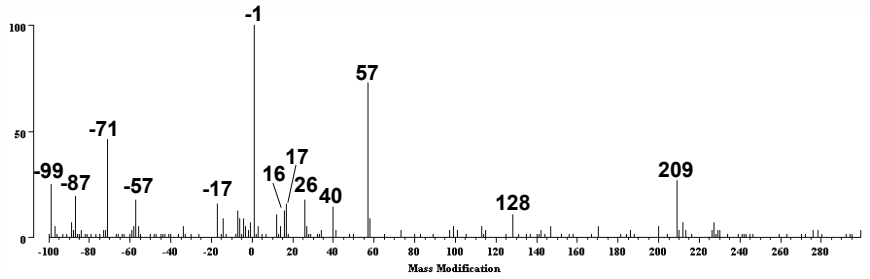
• No enzyme
• Non-Specific at N termini
• Non-Specific at C termini
} Unique to Protein Prospector
• Non-Specific at 1 termini (N or C but not both)
• Non-Specific at 2 termini (like No Enzyme but considers missed cleavages)

• This sample had a larger than usual number of peptide with non-specific cleavages but very few with missed cleavages.
  • Product of microwave-assisted digestion.

Type	1st Pep	2nd Pep	Total	%
0	669	43	712	57
N	246	8	254	20.4
C	220	7	227	18.2
2	53	2	55	4.4

Sometimes a second co-eluting peptide could be identified in a single MSMS spectrum.

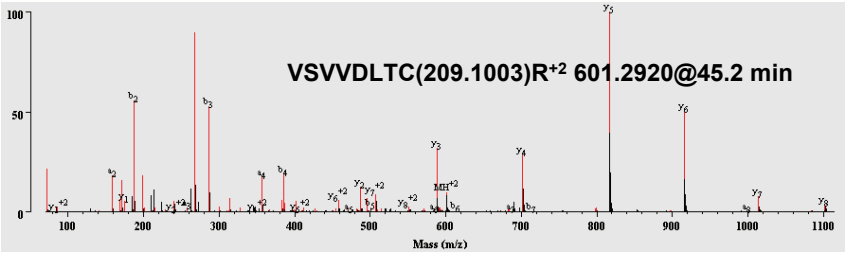
Mass Modifications



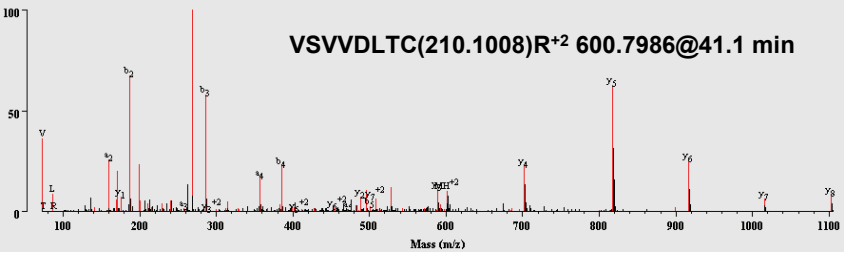
• Prospector can display a histogram showing the mass modifications found.

• Results can be sorted by mass modification.
  • If a modification is observed several times it is more likely to be real.
  • Helps identify novel modifications such as C(209).
    • This modification is formed by the alkylation of the Cys-SH with dithiothreitol, then alkylation of DTT on its other –SH group by iodoacetamide (Chalkley et al 2008).

Unexpected Modifications - C(209) and C(210)



An example of a spectra with the C(209) modification.

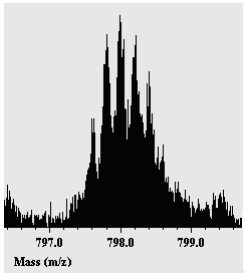


C(210) was assigned as Cys(Carboxymethyl-DTT)-derivative

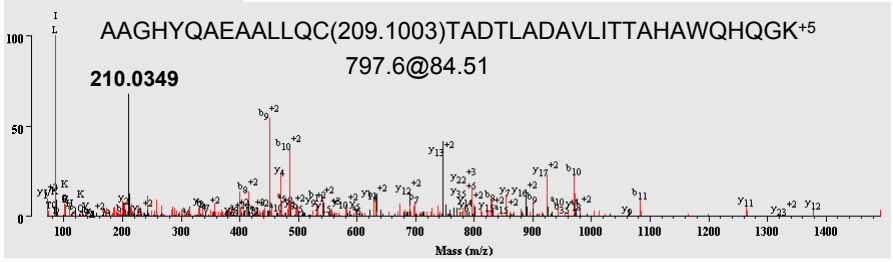
List of Modifications Found

- C(Carbamidomethyl), C(Dehydro)
- C(209), C(210) – see above
- C(12) – tentatively assigned as carbon incorporation similar to P(12)
- C(32) – tentatively assigned as persulfide
- C(56) – maybe Acrolein 56
- -17Da - ring formation from C(Carbamidomethyl) at N-terminus
- E(128) – possible insertion of Lys residue (see below)
- H(Methyl) - Tele-methylhistidine in Alpha skeletal muscle actin - listed in Swiss Prot
- K(Methyl), K(Dimethyl) - In trypsin
- K(Delta:H(6)C(6)O(1)) – maybe Acrolein 94
- M(Met-loss+Acetyl), M (Met-loss) – both from N-terminal processing
- M (Oxidation)
- N(Asn->Succinimide), Asn(Deamidated)
- P(-30) - Maybe Pro->Pyrrolidinone
- P(12) – See Nielsen 2006
- Q(Gln->pyro-Glu)
- +17 Neutral loss - Ammonium adduct
- +32 Neutral loss
- +41 Neutral loss – Acetonitrile adduct
- W(12) – tentatively assigned as carbon incorporation similar to P(12)
- W(Oxidation)

Incorrect Charge /Monoisotopic Mass

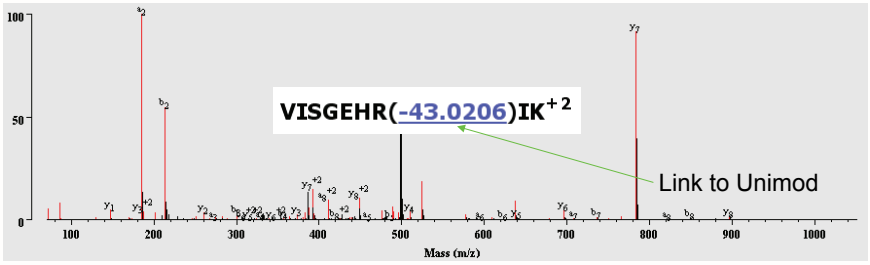


This 5+ peak had the charge and monoisotopic mass incorrectly assigned in the peak list. The charge and precursor m/z were determined by examining the precursor scan and the hit was found by using MS-Tag from File, manually specifying the charge.



Homologous Peptides

- Several spectra are of homologs of peptides found in the proteins identified in this dataset:
- Initially found by either:
  - mass modification search
  - homology tick box (amino acid substitution) option.
- Can be checked by searching a large database (NCBI/UniprotKB) using MS-Pattern or MS-Homology.



Homologous Peptides

PSI-MS Name	Interim name	Description	Monoisotopic mass	Average mass	Composition
	Asp->Ala	Asp->Ala substitution	-43.989829	-44.0095	C(-1) O(-2)
Arg->GluSA	Argglutamicsealde	Arginine oxidation to glutamic semialdehyde	-43.053433	-43.0711	H(-5) C(-1) N(-3) O
	Arg->Ile	Arg->Ile or Arg->Leu substitution	-43.017047	-43.0280	H(-1) N(-3)
	Val->Gly	Val->Gly substitution	-42.046950	-42.0797	H(-6) C(-3)
Arg->Orn	Arg2Orn	Ornithine from Arginine	-42.021798	-42.0400	H(-2) C(-1) N(-2)

Hits from Unimod for -43 Da

MS-Pattern

Database UniProt.20071010 | Digest No enzyme
DNA Frame Translation 3
Species All
Output Type HTML | Hits to file | Name lastres
[+] Pre-Search Parameters
Start Search | Sample ID (comment)
Reg Expression (Use CAPS) VISGEHLIK | Maximum Reported Hits 200
Max. # of Mismatched AA's 0 | Pre Search Only

MS-Pattern Search Results

Search completed. 11 sec elapsed. 0 sec remaining.

[+] Parameters

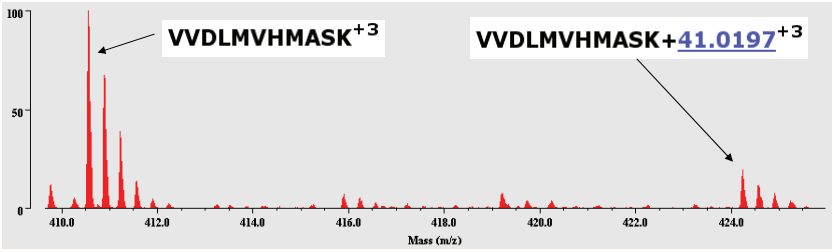
[+] Pre Search Results

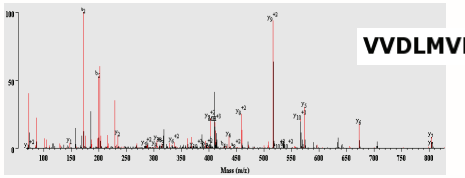
MS-Pattern search selects **1** entry.

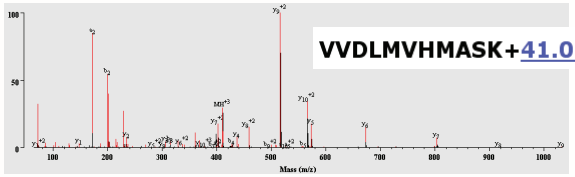
Matching Sequence	Peptide M+H	MS-Digest Index #	Protein MW (Da)/pI	Accession #	Species	Protein Name
(R)VISGEHLIK(A)	995.5884	3439688	58493/6.3	Q208A7	BACLI	Thermotolerant alpha-amylase

MS-Pattern identifies VISGEHLIK from Q208A7 from UniprotKB as a possible hit.

A Neutral Loss - Acetonitrile

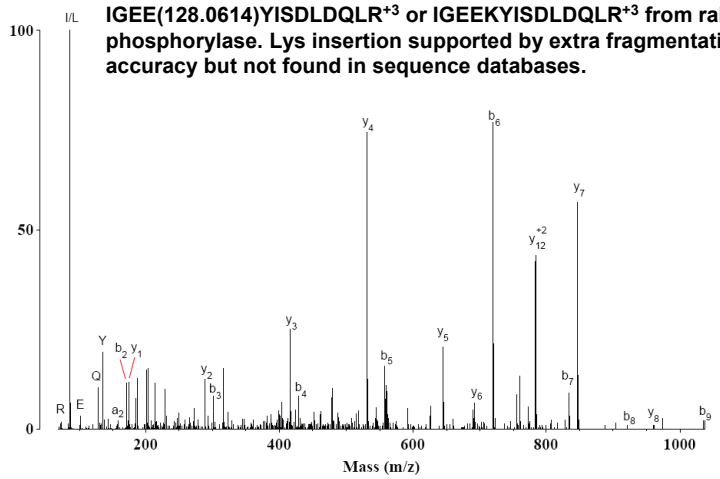






Inserted Amino Acid

The dataset contains one example of what is thought to be an inserted amino acid. This was found by the mass modification search.



IGEEKYISDLDQLR is a better match than IGEKEYISDLDQLR

Conclusions

- Programs in Protein Prospector allowed interpretation of many more (the majority) of the spectra:
- Regular database search identifies about 20% of spectra.
- Allowing for defined variable modifications and non-tryptic cleavages, Sequest identified about 25% of spectra (Klimek et al.)
- Improved Protein Prospector (with mass modification serching) along with manual inspection of the matches identified 64% of the spectra.
- If the spectra at the start and end of the run, where there are no peptides, are excluded 77% are interpreted

This software is freely available for use on the web at:

http://prospector2.ucsf.edu

Acknowledgements

Funding: NIH NCRR grant 001614 and the Vincent Coates Foundation.

References

1) Klimek, J., Eddes, J. S., Hohmann, L., Jackson, J., Peterson, A., Letarte, S., Gafken, P. R., Katz, J. E., Mallick, P., Lee, H., Schmidt, A., Ossola, R., Eng, J. K., Aebersold, R., and Martin, D. B., The Standard Protein Mix Database: A Diverse Data Set To Assist in the Production of Improved Peptide and Protein Identification Software Tools. *J Proteome Res*, 2007.

2) Baker, P. R., Medzihradszky, K. F., and Burlingame, A. L. Improved Methods for Comprehensive Sample Analysis Using Protein Prospector. *Proceedings of the 54th ASMS Conference*. 2006. Seattle, WA.

3) Schlosser, A. and Volkmer-Engert, R. Volatile polydimethylcyclosiloxanes in the ambient laboratory air identified as source of extreme background signals in nanoelectrospray mass spectrometry. *J. Mass Spectrom*., **38:** 523-525 (2003)

4) Pickering, G. R., Olliff, C. J. and Rutt, K. J. The mass spectrometric behaviour of dimethylcyclosiloxanes. *Organic Mass Spectrometry.,* **10:** 1035-1045 (1975)

5) Chalkley, R. J., Baker, P. R., Medzihradszky, K. F., Lynn, A. J. and Burlingame, A. L. In-depth Analysis of Tandem Mass Spectrometry Data from Disparate Instrument Types. *Submitted* (2008)

6) Baker P. R., Chalkley R. J., Medzihradszky K. F. and Burlingame A. L. Discovery of Unanticipated Modifications using Protein Prospector, *Proceedings of the 55th ASMS Conference*. 2007 Indianapolis, IA

7) Nielsen, M. L., Savitski, M. M. and Zubarev, R. A. Discovery and Unexpected Modifications by Modificomb: the Case of Glu-Methylation in Human Cells. *Proceedings of the 54th ASMS Conference*. 2006. Seattle, WA.