The Pitfalls of Peaklist Generation Software Performance on Database Searches

Aenoch J. Lynn, Robert J. Chalkley, Peter R. Baker, Katalin F. Medzihradszky, Shenheng Guan, and A.L. Burlingame

Department of Pharmaceutical Chemistry
University of California, San Francisco
San Francisco, California

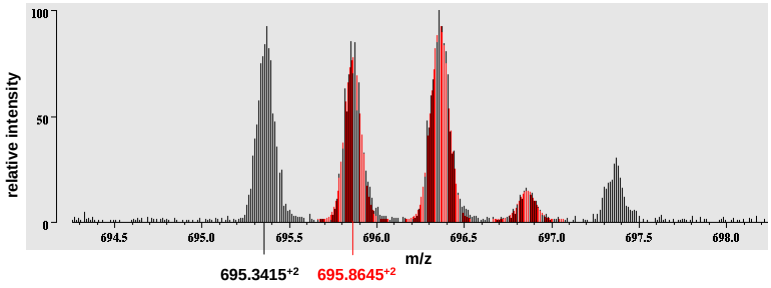
There are Many Steps Where Peaklist Generation Can Go Wrong

- Determination of precursor ion monoisotopic mass and charge
- Determination of fragment ion monoisotopic mass and charge
- Removal of the isotope peaks (deisotoping)
- Adjusting monoisotopic intensities for the removal of isotope peaks
- Summation of spectra with the same precursor m/z
- Duplication of peaklists and assignment of charge states when the precursor charge is unknown

Here we show some of the possible errors in generated peaklists and compare several methods to generating peaklists.

An Example of an Incorrect Precursor Monoisotopic Peak Selection

Precursor scan from a QStar wiff file.



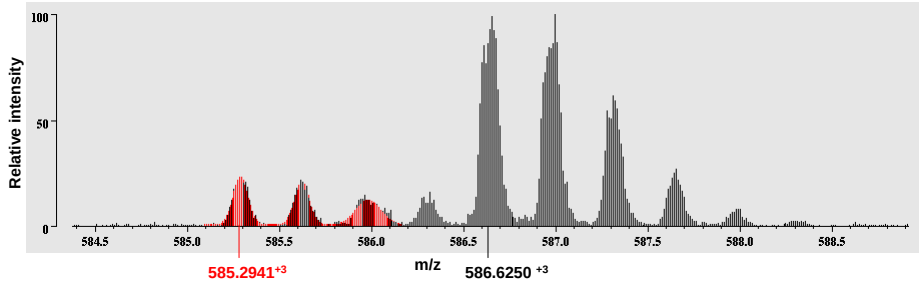
Centroiding program (Mascot.dll in AnalystQS 2.0) assigned 695.8645 as the precursor m/z, but this is the wrong precursor mass.

695.3415^{+2} – "Correct" monoisotopic m/z as determined from a mass modification search in Protein Prospector v5.0.1. Using this precursor mass resulted in a Protein Prospector score of 42.5 for the MSMS from this precursor.

695.8645^{+2} – Peaklist monoisotopic m/z. Incorrect isotope distribution as extracted from the QStar file during peaklist generation. Database search using this precursor mass resulted in a Protein Prospector score of 11.4 for the best matching sequence.

An Example of Co-eluting Precursor Ions

Precursor scan from a QStar wiff file.



Centroiding program (Mascot.dll in AnalystQS 2.0) failed to assign precursor mass of a co-eluting ion with a strong signal.

585.2941^{+3} – Precursor m/z from generated peaklist. Probably the m/z that was targeted by the instrument for MSMS. Protein Prospector score of –1 using this precursor m/z.

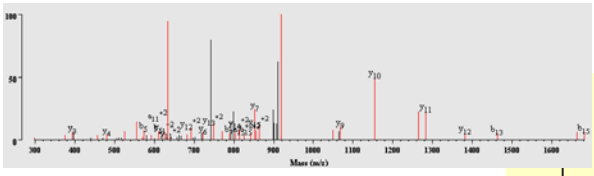
586.6250 $^{+3}$ – Precursor m/z of a co-eluting ion with an overwhelming ion signal. In a Protein Prospector mass modification search this resulted in a -4 neutral loss on 585.2941^{+3}. The peptide sequence had a Protein Prospector score of 44.8.

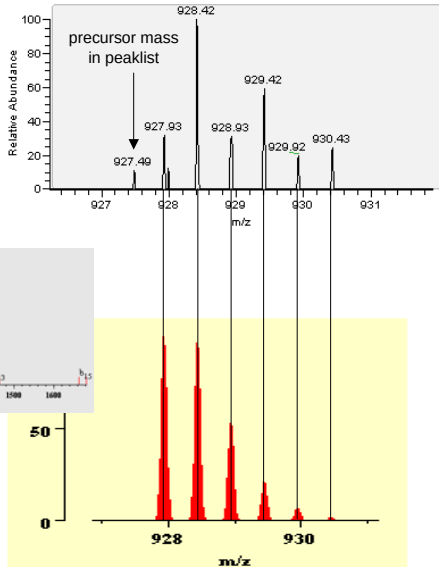
Which Peak is the Monoisotopic Peak?

It's not easy determining the monoisotopic peak.

A precursor scan taken by an LTQOrbitrap shows a complicated series of peaks. Any software is going to have a hard time determining which is the monoiostopic peak.

When the MSMS peaklist is searched using the mass modification search in Protein Prospector, the best scoring peptide has a precursor mass of 927.93 m/z whereas the peaklist had 927.49. Red peaks are matched sequence or loss ions, black are unmatched.





Theoretical isotope distribution for the ion 927.93^{+2} aligned with the experimental data. They are dissimilar.

Comparison of Software Approaches

To illustrate problems and mistakes peaklist generation software makes:

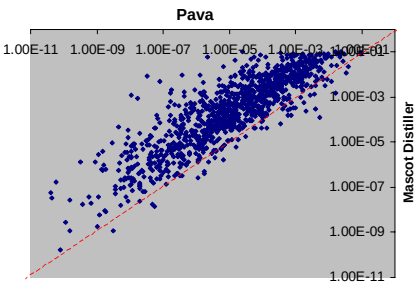
- Use of three different software for analysis of LTQ MSMS data:
  - Mascot Distiller (Matrix Science)
    - Attempts peak detection from profile raw data, then de-isotopes data.
  - Pava (UCSF in-house software)
    - Streamlined peak detection, de-isotoping
  - ReAdW (opensource program from Seattle Proteome Center)
    - Raw data with no post-processing
- Peaklists were searched using Protein Prospector
  - New version 5.0 with batch MSMS searching released.
  - Available for public use.
    - http://prospector2.ucsf.edu/

If Data are Acquired in Centroid Mode then the Mascot Distiller Generates Poor Peaklists

The Mascot Distiller converts centroid data into profile data, then centroids the profile data.

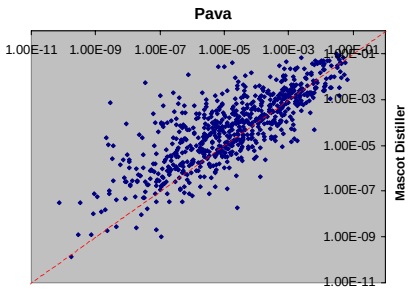
- Peaklists were generated by UCSF's Pava software and Mascot Distiller from LTQOrbitrap RAW data files and searched with Protein Prospector v5.0.1.
- For data saved in centroid mode, a plot of peptide expectation values shows that peptide assignments for the Mascot Distiller peaklist is significantly worse than the scores for the peaklist generated by UCSF's Pava.
- For data saved in profile mode, a plot of peptide expectation values shows similar scores for UCSF's Pava and the Mascot Distiler, but a larger variance.
- The Mascot Distiller has many parameters that can be changed; for this study the default LCQ parameters were used.

Relative Performance for Centroid and Raw Data for UCSF's Pava and the Mascot Distiller



Data saved in centroid mode

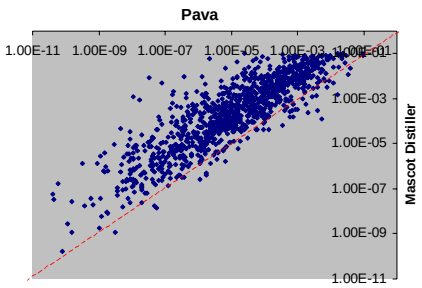
On centroided data, the Mascot Distiller generated peaklists have peptide scores systematically worse than UCSF's Pava generated peaklists.

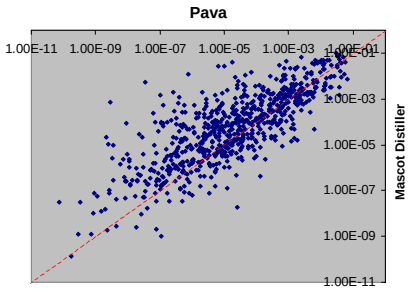


Data saved in profile mode

On profile data, the Mascot Distiller on average performs as well as UCSF's Pava, but the scores have a larger variance.

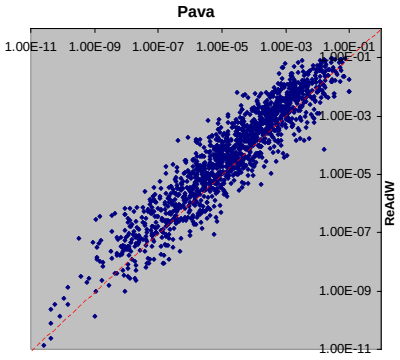
Relative Performance for Centroid and Raw Data for UCSF's Pava and the Mascot Distiller





Effect of Peak Detection and De-isotoping

Comparing the expectation values from database searches with peaklists generated by UCSF's Pava and ReAdW.



•Pava (with peak detection and de-isotoping) gives better expectation values than ReAdW peaklist (raw centroided data).

Unique Peptides Found with Each Software

Protein Prospector search results – peptides found using peaklists generated by the Mascot Distiller and UCSF's Pava can differ.

24 **Acc. #:** P08551 **Gene:** NFL_MOUSE **Species:** MOUSE **Name:** Neurofilament light polypeptide

Protein MW: 61508.9 **Protein pI:** 4.6 **Protein Length:** 543

Search Name	Num Unique	% Cov
Mascot Distiller generated peaklist	16	33.7
UCSF's Pava generated peaklist	18	34.4

Mascot Distiller generated peaklist

m/z	Peptide	Score	Expect
1090.5300	Q(Gln->pyro-Glu)LQELEDKQNADISAMQDTINKLENELR	42.4	1.2e-7
768.3768	QNADISAMQDTINKLENELR	35.5	6.5e-6
761.9084	IDSLMDEIAFLKK	30.1	1.7e-5
856.4293	Q(Gln->pyro-Glu)ALQGEREGLEETLR	28.2	6.1e-5
763.4134	YLKEYQDLLNVK	31.2	4.3e-6
1143.5497	Q(Gln->pyro-Glu)NADISAMQDTINKLENELR	28.3	5.4e-6
1057.1872	RSYSSSSGSLMPSLENLDLSQVAAISNDLK	24.2	6.1e-4
517.6124	RIDSLMDEIAFLK	37.0	1.6e-5
854.3987	SAYSGLQSSSYLMSAR	23.4	6.2e-4
531.2559	LAAEDATNEK	25.2	8.2e-5
923.9109	M(Met-loss+Acetyl)SSFGYDPYFSTSYKR	16.9	1.0e-4
496.2485	TLEIEAC(Carbamidomethyl)R	20.0	0.0015
722.3222	ESEEEEKKEESAGEEQVAK	27.6	1.4e-4
705.8608	IDSLM(Oxidation)DEIAFLK	21.6	2.1e-4
435.2304	FASFIER	16.6	0.0010
512.7485	YEEEVLSR	29.5	0.0022

UCSF's Pava generated peaklist

m/z	Peptide	Score	Expect
1090.5347	Q(Gln->pyro-Glu)LQELEDKQNADISAMQDTINKLENELR	50.8	1.4e-9
768.3793	QNADISAMQDTINKLENELR	37.8	2.2e-7
761.9117	IDSLMDEIAFLKK	27.4	2.7e-6
856.4326	Q(Gln->pyro-Glu)ALQGEREGLEETLR	28.8	3.2e-6
763.4160	YLKEYQDLLNVK	38.2	1.1e-5
1143.5485	Q(Gln->pyro-Glu)NADISAMQDTINKLENELR	28.2	1.2e-5
1057.1919	RSYSSSSGSLMPSLENLDLSQVAAISNDLK	27.6	5.4e-6
775.9145	RIDSLMDEIAFLK	34.7	6.2e-6
854.4016	SAYSGLQSSSYLMSAR	28.1	2.7e-5
531.2568	LAAEDATNEK	17.2	6.1e-5
496.2471	TLEIEAC(Carbamidomethyl)R	25.1	1.1e-4
722.3253	ESEEEEKKEESAGEEQVAK	24.8	1.7e-4
697.3766	MALDIEIAAYRK	16.6	3.4e-4
435.2304	FASFIER	20.7	6.7e-4
862.3990	SAYSGLQSSSYLM(Oxidation)SAR	16.9	9.0e-4
512.7504	YEEEVLSR	30.9	0.0014
473.7441	LLEGEETR	19.6	0.0018
822.4114	QLQELEDKQNADISAMQDTINKLENELR	19.9	0.0034

Conclusions

- Some level of processing of the data (as in UCSF's Pava) is beneficial for improving data analysis.
- The Mascot Distiller is much more effective when operating on profile data rather than centroid data.
- Pava is effective on both profile and centroided data.
- The various centroiding programs each have their own strengths and weaknesses.
  - Using peaklists generated by different programs allows the identification of more peptides.
- A manuscript on UCSF's Pava is in preparation.

This work was supported by NIH NCRR 01614 and the Vincent J. Coates Foundation.