Mascot DIA introduces probabilistic scoring to spectrum-centric analysis of DIA data

Ville Koskinen[1], Patrick Emery[1], Richard Jacob[2]*

** villek@matrixscience.com [1]Matrix Science Ltd, London, UK [2]Matrix Science Inc, Boston, MA, USA*

https://www.matrixscience.com/contact.html

1. Spectrum-centric analysis

Most software packages for data-independent acquisition (DIA) are peptide-centric. They depend on a precomputed spectral library with accurate fragment intensity predictions, limiting the choice of instrument, proteome, enzyme and variable modifications.

We have developed a new probabilistic scoring model for identifying peptides from convoluted DIA MS/MS spectra. The model is independent of machine learning and overcomes peptide-centric limitations.

2. Scoring model

Given an MS/MS spectrum and an observed precursor mass, Mascot selects peptide sequences within mass tolerance and fragments them *in silico*. Chimeric precursors are matched independently and concurrently. MS/MS peaks are selected in order of decreasing intensity and matched to fragments with the null hypothesis (H_0): *peak matches are independent and identically distributed, occurring at a constant rate.* The (multiple-testing corrected) p-value reflects how much the peptide match diverges from randomness.

3. Evaluation data

The only real difference between DDA and DIA spectra is the amount of 'noise' peaks. Thus, we compared the performance to the Mascot 3.1 scoring model, holding all other factors constant. Chimeric precursor detection and peak picking were done using Mascot Distiller 3.0 beta (see also WP 464).

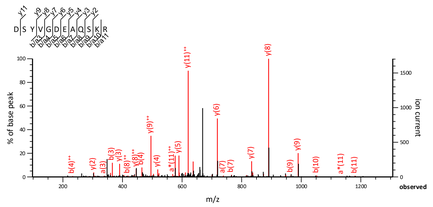
A: CPTAC study 6 spike-in (2010). Lots of background and chemical noise. DDA, 0.6Da tol., a/b/y/-H2O/-NH3 ions.

B: PXD024275 (2021). Phospho-enriched TMTpro 18-plex. DDA, 0.02Da tol., b/y/-H2O/-NH3 ions plus phospho NL.

C: PXD028735 (2022). Thermo AIF (DIA), 8Th isolation window, 20ppm tol., b/y/-H2O/-NH3 ions.

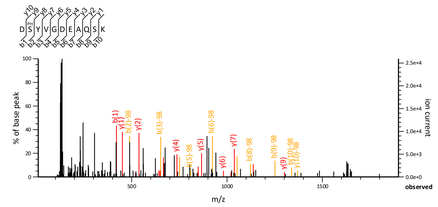
4. Results and conclusions

A



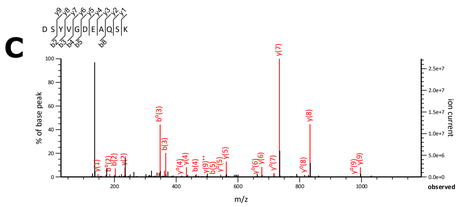
Model	PSMs	Sequences
3.1 DDA	2648	2403
3.2 DIA	2610	2357
	-1.4%	-1.9%

B



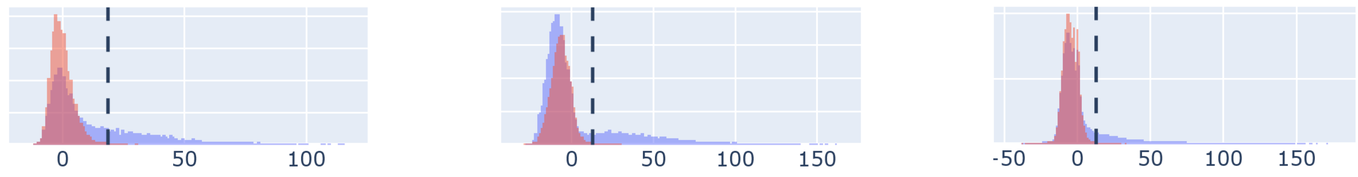
Model	PSMs	Sequences
3.1 DDA	6051	4521
3.2 DIA	6704	4996
	+10.8%	+10.5%

C

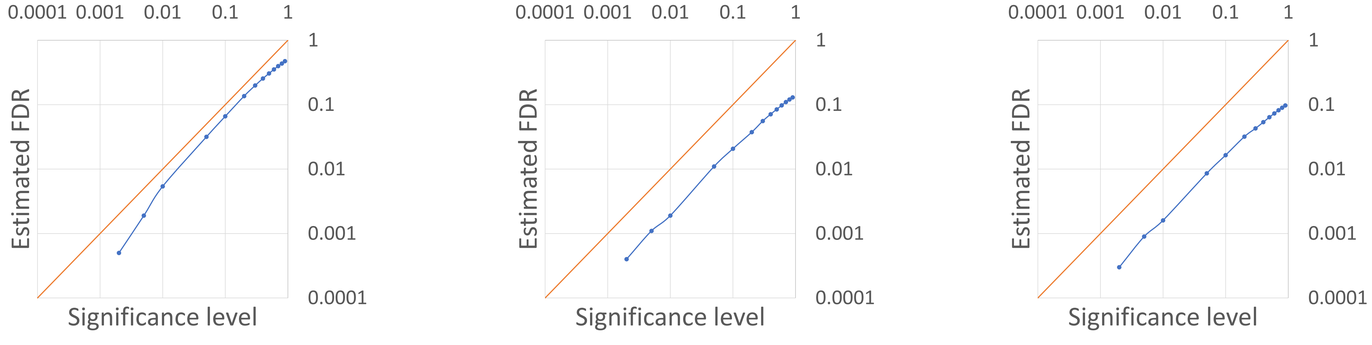


Model	PSMs	Sequences
3.1 DDA	64683	8762
3.2 DIA	112397	14570
	+73.8%	+66.3%

Sensitivity at 1% FDR is greatly improved with DIA. The new scoring model is *noise resistant* regardless of the type of interfering peaks and adapts to very different fragmentation patterns.



Decoy matches accurately model incorrect target matches. Histogram of $-10\log_{10} E$, where E is multiple-testing corrected p-value; target is blue, decoy red; dashed line is 1% PSM FDR threshold.



Significance level α controls Type I (FP) errors – unique feature of probabilistic scoring. 'Null' p-values under H_0 have Uniform(0,1) distribution. If all null hypotheses are true, then the proportion of matches with $p \leq \alpha$ (e.g. 0.01) is at most α (e.g. 1%). False discovery rate $\mathrm{FP}/(\mathrm{TP}+\mathrm{FP})$ never exceeds α.

COI: Ville Koskinen and Patrick Emery are directors and minority shareholders of Matrix Science Ltd.