Bioinformatics Solutions Inc.

Uncover the hidden peptidome/proteome with sequence variants and novel peptides from DIA data

Baozhen Shan, PhD, CEO
Natalie Korkola, PhD, Application Scientist

Bioinformatics Solutions Inc., Waterloo, Canada

Abstract

Data-independent acquisition (DIA) mass spectrometry (MS) is a useful acquisition method for increased reproducibility and depth of proteome coverage. In particular, there is potential to uncover low-abundance peptides from the "dark proteome" that may be missed in traditional DDA methods. However, it is important to employ a data analysis method that can identify proteins and peptides which may not be included in a traditional database. PEAKS 13 provides a complete DIA analysis solution for accurate and sensitive identification and quantification. The database search method can be integrated with sequence variant analysis and *de novo* sequencing to ensure that peptides that may not be included in a database are reported.

Introduction

Numerous studies have demonstrated the existence of a "dark proteome", consisting of proteins critical for biological processes but not included in widely used gene catalogs. Data-independent acquisition (DIA) mass spectrometry (MS) has emerged as a powerful technology for high-throughput, accurate, and reproducible quantitative proteomics. In addition, DIA provides the capability to identify low abundance peptides that may be missed in traditional DDA methods. For DIA data analysis, major strategies can be classified into spectrum reconstruction, sequence-based search, library-based search, *de novo* sequencing, and sequencing independent approaches. [1] Here, we demonstrate PEAKS DIA, which integrates sequence-based search and *de novo* sequencing, to identify database peptides, homolog peptides, and novel peptides. This novel approach has the potential to shed some light on the dark proteome.

In this Application Note, we demonstrate how PEAKS DIA can be used to accurately identify sequence variants and novel peptides from a human dataset. Sequence variants from isoforms which are not included in the database can be identified. A simulation test was also conducted where an incorrect mouse database was used to search the human dataset. It was demonstrated that PEAKS could identify the correct human peptides through sequence variants compared to the mouse database and through the identification of novel peptides. This shows that PEAKS could be used to accurately identify peptides from the correct species even in cases where these sequences were not present in the database.

Methods

The PEAKS 13 DIA algorithm contains three steps.

1. Given a fasta sequence database and DIA data, it first performs direct database search. Each identified precursor is associated with a peptide sequence in the database and a score. Also, each positional amino acid confidence is recorded.

2. Second, spectrum reconstruction and *de novo* sequencing is performed for each precursor. The corresponding *de novo* peptide sequence and its positional amino acid confidence score are recorded.

3. Finally, the reconstructed peptide sequence of each precursor is derived from database peptide and *de novo* peptide. There, the sum of the distance between reconstructed sequence and database peptide and the distance between reconstructed sequence and *de novo* peptide is minimized in the algorithm.

A DIA data set from the ABRF Proteomics Research Group (PRG) 2017/2018 study was used [2]. The data were acquired from HeLa digests with Orbitrap

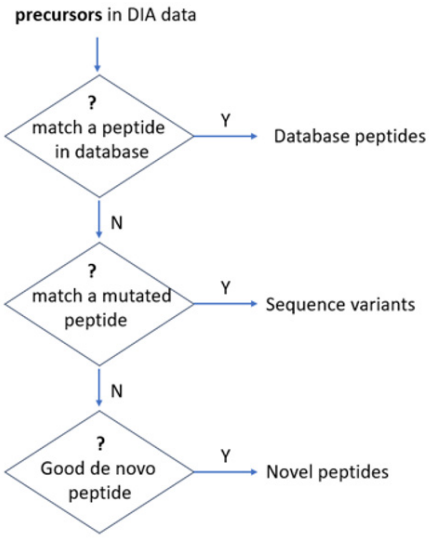


Figure 1. The PEAKS 13 DIA Proteome workflow steps used in this Application Note.

The data were searched in PEAKS 13 using the DIA Proteome workflow, which includes library search, database search, sequence variant analysis, *de novo* sequencing, label-free quantification, and quality control options. In this application note, the steps used were a direct database search (skipping the optional library step), sequence variant analysis, and *de novo* sequencing (Figure 1). The search parameters are shown in Table 1. The data were searched against a human database and a mouse database in separate analyses.

Results

PEAKS 13 DIA reveals isoform sequences of human proteins not included in the database and novel peptides that map to human proteins

The DIA database search result includes a protein coverage view, showing database peptides, sequence variants, and novel peptide tags which map to proteins in the provided database. The coverage map of Nesprin-2 is shown in Figure 2. The peptide QATSDVQESTQESAT(sub A) VEK was identified, which contains a sequence variant. Highlighting over the peptide containing the sequence brings up an alignment which shows the differences between the *de novo* sequence, the database peptide, and the resulting real peptide reported by PEAKS. The PEAKS algorithm uses both *de novo* sequencing errors and sequence variants to explain the difference between the database sequence and the *de novo* sequenced peptide. In this case, "QA" was identified as a sequencing error due to low direct fragmentation evidence at the N-terminal end of the peptide. Good supporting fragmentation with high ion intensity supports the assignment of the T→A substitution.

Table 1. PEAKS Studio 13 DIA database search parameters for a search against (A) human database and (B) mouse database.

Analysis parameters	Settings
Precursor mass tolerance	Auto Detected
Fragment mass tolerance	Auto Detected
Enzyme	Trypsin (Semi-specific)
Fixed PTMs	Carbamidomethylation (C)
Variable PTMs (max 1 per peptide)	Deamidation (NQ) Oxidation (M)
Database	(A) Homo Sapiens (20 199 sequences) (B) Mus musculus (16 718 sequences)
Sequence Variants Filter	-10LgP ≥ 20
de novo Filter	ALC ≥ 80%
Protein Group FDR	1.0%
Precursor FDR	1.0%

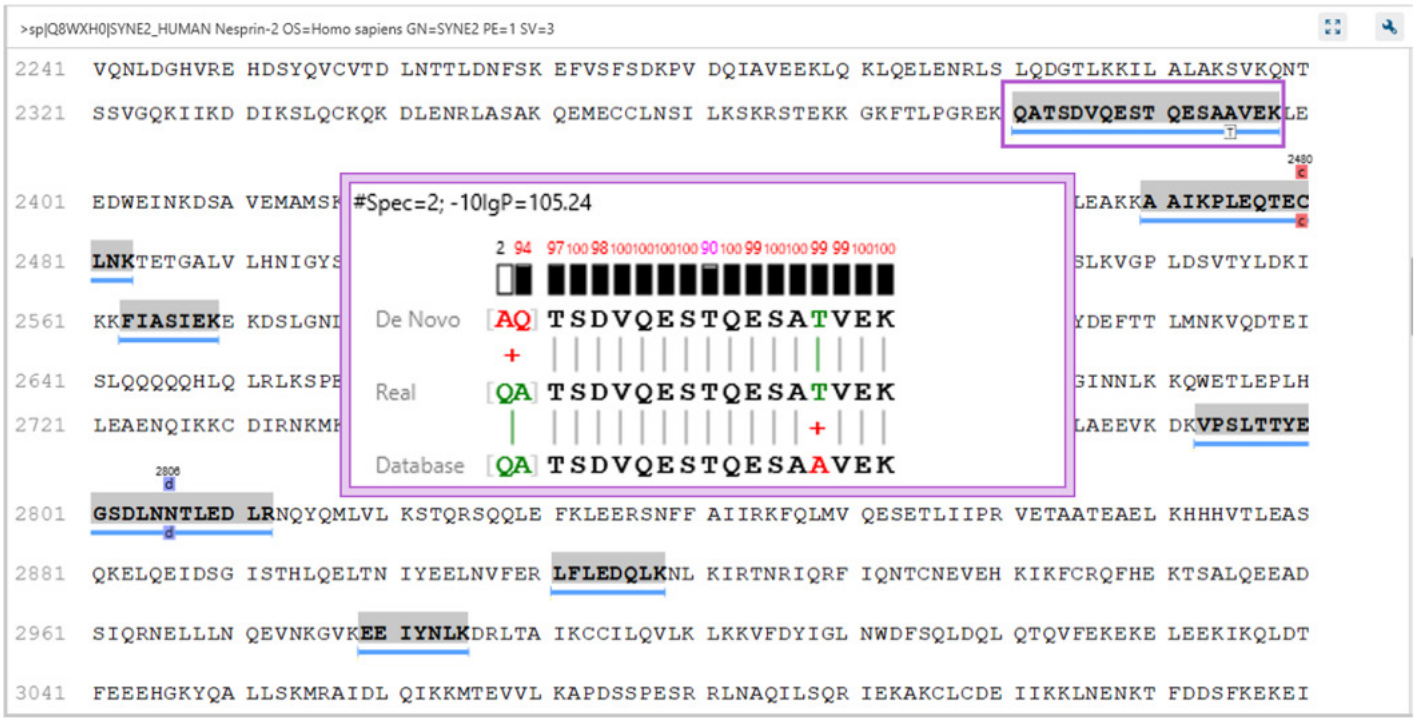


Figure 2. Protein coverage view of Nesprin-2. The peptide QATSDVQESTQESAT(sub A)VEK contains an A→T substitution which was detected by PEAKS. The *de novo* sequenced peptide, the reported peptide, and the database sequence are aligned with amino acid level confidence for easy result validation.

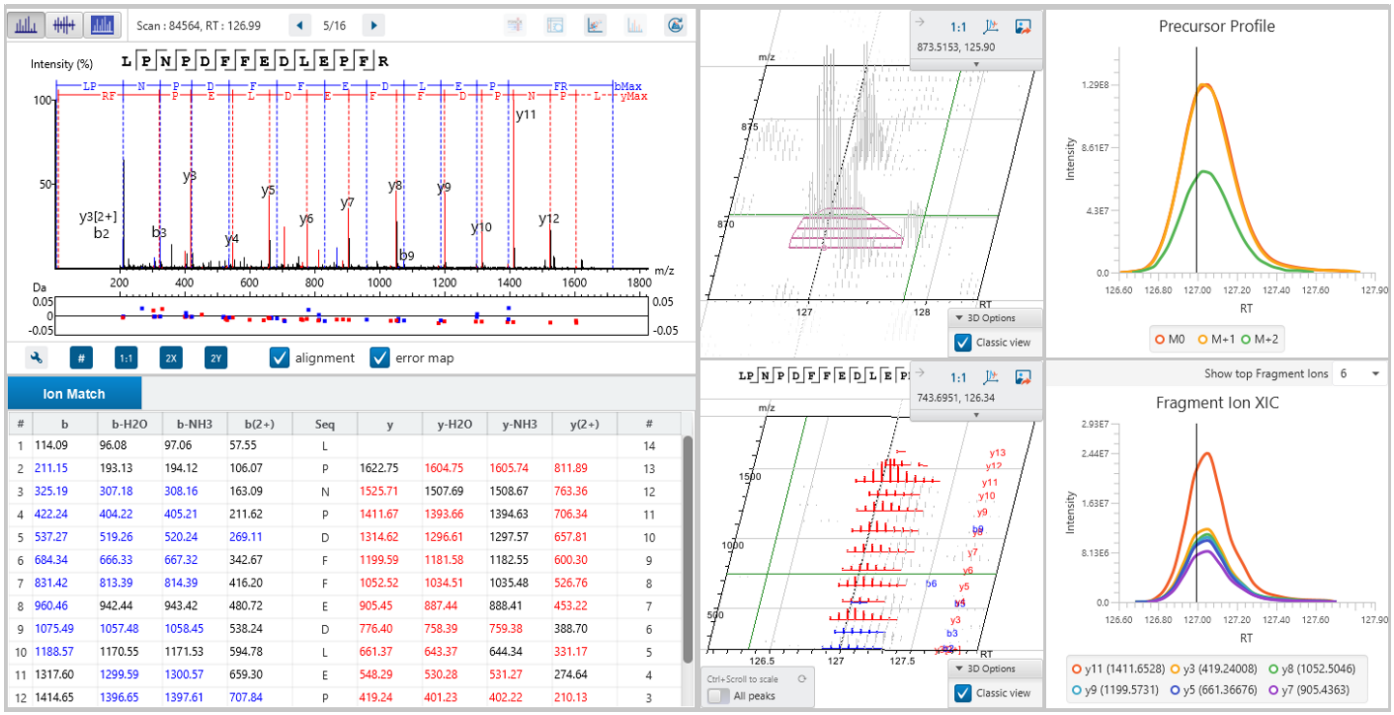
A protein blast analysis revealed that this sequence variant is from human nesprin-2 isoform X1. The knowledge that this sequence variant belongs to a human protein increases the result confidence, as human proteins are expected to be found in these samples.

Individual result tabs for database peptides, sequence variants, and novel peptides are also provided. For each result, the supporting MS2 spectra as well as LCMS maps and profiles at the precursor and fragment ion levels are provided for easy validation of the result. A single peptide identification can be quickly linked back to the full LCMS data view as an easy way to view surrounding identifications.

The identified novel peptide, LPNPDFFEDLEPFR is shown in Figure 3. The peptide score is 99%, an assignment which is supported by sequentially assigned fragment ions. High-quality XIC signals which overlap at the precursor and fragment ion level provide a visual method of validating the result. This example shows that the novel peptides identified make biological sense.

The protein blast result shows that LPNPDFFEDLEPFR is a tryptic peptide from human calnexin isoform a. The assignment of the novel peptide in PEAKS makes biological sense for the experiment.

This demonstrates result accuracy because the novel peptide found was expected for human samples. This demonstrates that PEAKS can detect novel peptides from the correct organism when the sequences are not found in the database.



#	b	b-H2O	b-NH3	b(2+)	Seq	y	y-H2O	y-NH3	y(2+)	#
1	114.09	96.08	97.06	57.55	L					14
2	211.15	193.13	194.12	106.07	P	1622.75	1604.75	1605.74	811.89	13
3	325.19	307.18	308.16	163.09	N	1525.71	1507.69	1508.67	763.36	12
4	422.24	404.22	405.21	211.62	P	1411.67	1393.66	1394.63	706.34	11
5	537.27	519.26	520.24	269.11	D	1314.62	1296.61	1297.57	657.81	10
6	684.34	666.33	667.32	342.67	F	1199.59	1181.58	1182.55	600.30	9
7	831.42	813.39	814.39	416.20	F	1052.52	1034.51	1035.48	526.76	8
8	960.46	942.44	943.42	480.72	E	905.45	887.44	888.41	453.22	7
9	1075.49	1057.48	1058.45	538.24	D	776.40	758.39	759.38	388.70	6
10	1188.57	1170.55	1171.53	594.78	L	661.37	643.37	644.34	331.17	5
11	1317.60	1299.59	1300.57	659.30	E	548.29	530.28	531.27	274.64	4
12	1414.65	1396.65	1397.61	707.84	P	419.24	401.23	402.22	210.13	3

Figure 3. The Novel Peptide result view for the peptide LPNPDFFEDLEPFR.

PEAKS identifies sequence variants and novel peptides corresponding to the correct species when an incorrect database is used

A simulation test was conducted. Rather than searching the human data using the human database, the dataset was searched against a mouse database. The peptide identifications were summarized in Table 2.

Table 2. Comparison of peptides identifications between searching human and mouse databases

	# DB peptides	# Variants	# Novel peptides
Human	47900	127	2181
Mouse	23262	611	2851

It can be seen from Table 2 that using the correct human database to search the human dataset resulted in approximately twice the number of database peptide identifications compared to when an incorrect mouse database was used. Conversely, 4.8 times more sequence variants and 1.3 times more novel peptides were identified when the data was searched against the mouse database compared to the human.

A detailed comparison of the overlap in peptide identifications between searching the human and mouse databases is shown in Figure 4.

Approximately 67% of the sequence variants identified from searching the mouse database are peptides that were identified from searching the human database.

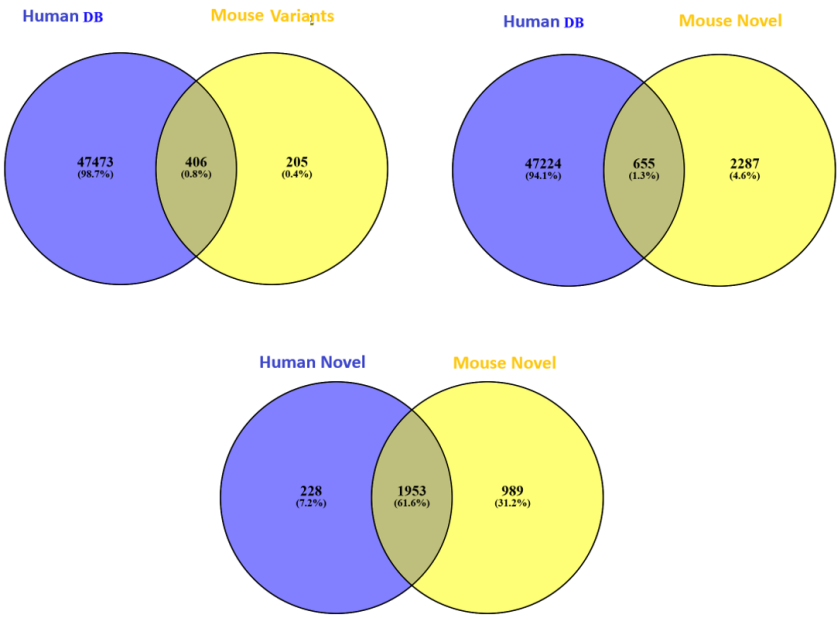


Figure 4. Venn diagrams showing the overlap of peptides identified from searches of a human dataset against a human database and mouse database.

Approximately 23% of the novel peptides identified from the search against the mouse database are peptides which were identified in the human database. Approximately 90% of the novel peptides identified from searching the human database are the same novel peptides as those identified from searching the mouse database.

These results demonstrate the high sensitivity and accuracy of the PEAKS DIA database search, sequence variant analysis, and *de novo* sequencing. When the incorrect mouse database was used to search the data, The PEAKS DIA algorithm could correctly identify the human peptides from the sequence variation between the two species and report novel peptides that could not be mapped to the mouse database but are known to be human peptides. This shows that PEAKS DIA is a valuable tool in cases where a database is incomplete or the sample is partially unknown.

PEAKS DIA results can also be used to assess the appropriateness of the database selected for the analysis. If a high proportion of the result comes from sequence variants or novel peptides, it may indicate that a different database should be used.

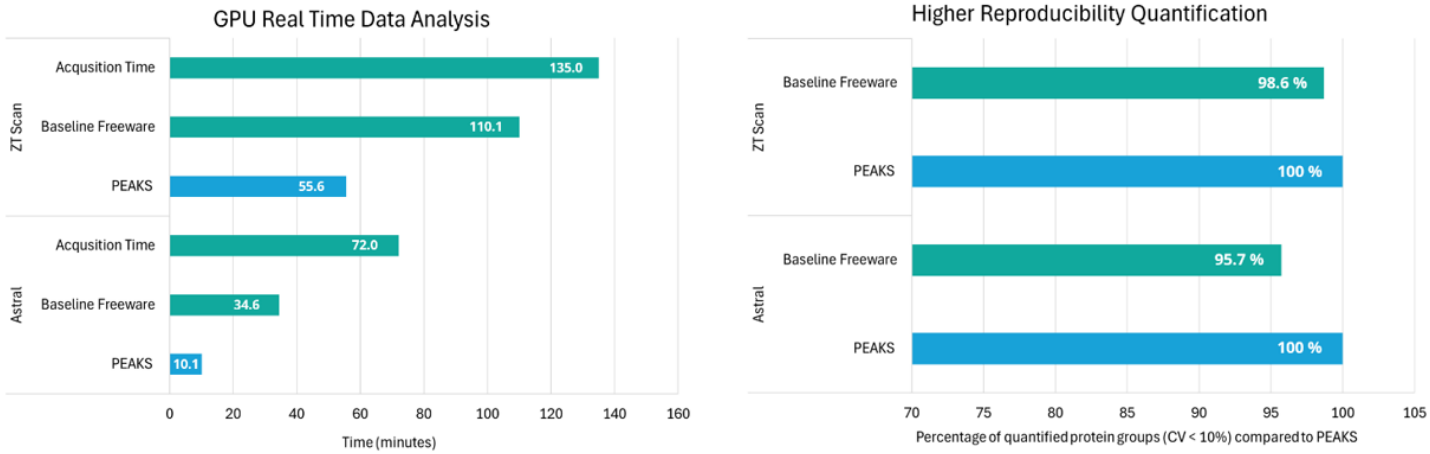


Figure 5. Benchmarking of PEAKS 13 DIA speed and reproducibility for Astral and ZT Scan compared to Freeware.

PEAKS 13 DIA provides faster analysis times and higher reproducibility

Figure 5 shows the benchmarking results of PEAKS 13 DIA database search compared to Freeware alternatives for two instruments. PEAKS 13 analysis is approximately 2 times faster for ZT Scan and approximately 3 times faster for Astral compared to Freeware. For both ZT Scan and Astral, 100% of protein groups quantified by PEAKS have a CV below 10%, while the freeware has 98.6% and 95.7% of quantified protein groups below a CV of 100%, respectively.

Conclusions

PEAKS 13 DIA provides a fast, accurate, and sensitive method for identification and quantification of DIA proteomics data. The inclusion of sequence variants and novel peptide searches into the algorithm makes PEAKS a useful tool to uncover the “Dark Proteome” of proteins and peptides that are not found in traditional database searches.

References

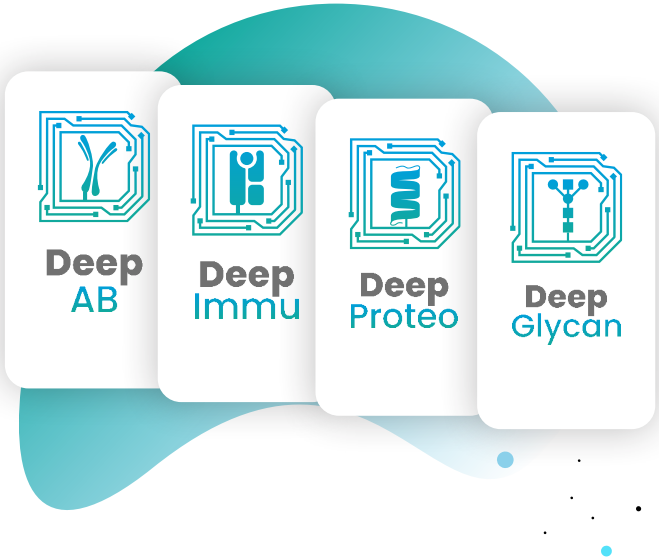
[1] Lou R, Shui, W. Acquisition and Analysis of DIA-Based Proteomic Data: A Comprehensive Survey in 2023, MCP, Volume 23, Issue 2, 2024, 100712, ISSN 1535-9476, DOI: 10.1016/j.mcpro.2024.100712.

[2] J Biomol Tech. 2019 Dec;30(Suppl):S46

FULL BSI SOFTWARE AND SERVICE SUITE

PEAKS®

DeepSuite



www.bioinfor.com
peaks@bioinfor.com
+1 855-885-8288
DeepProteomics
bsi-peaks
bsi-peaks
bsi_peaks
Bioinformatics Solutions Inc.

Bioinformatics Solutions, Inc.

140 Columbia St, Suite 202
Waterloo, Ontario N2L 3K8
Canada

Tel: (519) 885-8288
Fax: (519) 885-9075

sales@bioinfor.com
www.bioinfor.com