Improvements to ProSightPD nodes in the Thermo Scientific Proteome Discoverer Software Framework

David M. Horn,¹ Tara L. Schroeder,¹ Ioanna Ntai,¹ Richard D. Leduc,² Ryan T. Fellers,² and Neil L. Kelleher², ¹Thermo Fisher Scientific, San Jose, CA, ²Northwestern University, Evanston, IL

ABSTRACT

Purpose: Here we demonstrate the use of the ProSightPD[™] nodes for the Thermo Scientific[™] Proteome Discoverer[™] software framework to analyze complex top down proteomics data. We also demonstrate how sliding window deconvolution results from Thermo Scientific[™] Biopharma Finder[™] can be used to determine differentially expressed proteoforms.

Methods: For this work, Study 1 data from Ntai *et al*¹ were used for the analysis. Both the "High/High" and "Low/High" GELFrEE fractions were analyzed using ProSightPD Base and ProSightPD High Mass, respectively. The High/High data were analyzed using a five-step search to maximize the number of identifications, while the Low/High data were analyzed using a three-step search. PrSMs were filtered by a minimum C score of 3. Sliding window deconvolution in the Biopharma Finder software was used to process the High/High data using Xtract deconvolution. The quantitative results were integrated with the filtered human PrSMs from ProSightPD via a Perl script. Normalization and p-value calculations were produced using InfernoRDN and Microsoft® Excel.

Results: The High/High runs identified 2,374 proteoforms total with 735 producing a C-score of 3 or better. The Low/High runs identified 254 proteoforms with 38 proteoforms having a C-score of 3 or better. Many of the proteoforms that were identified had masses greater than 40 kDa. Integration of the sliding window deconvolution results with the High/High data produced over 400 proteoforms that differ significantly between the WHIM2 and WHIM16 samples.

INTRODUCTION

ProSightPC[™] is the leading search engine for top-down proteomics because of its sophisticated tools for proteoform identification. ProSightPC has several search modes for identification of intact and truncated proteins and uses the information annotated in UniProt to identify heavily modified and processed proteins. The user interface of ProSightPC is well suited for comprehensive characterization of single datasets, but it can be challenging to perform multi-stage searches against large datasets with many data files.

The ProSightPD nodes for the Proteome Discoverer software framework were created to analyze such complex datasets. The workflow layout makes it straightforward to create multi-step searches and the results interpretation tools are well suited for complex data. In the ProSightPD 1.1 release, additional nodes for sequence tag prefiltering prior to absolute mass or biomarker searches were added as well as the C-score² for confidence in proteoform characterization. Here we present further improvements to the ProSightPD nodes, including new result tables better suited for top down proteomics as well as integration of a fragment map similar to ProSight Lite.³

MATERIALS AND METHODS

High/High data analysis

The study 1 top down dataset from Ntai *et* al was downloaded from the CPTAC website. The multidimensional fractions for the High/High and Low/High data for the WHIM2 and WHIM16 samples were loaded into the Proteome Discoverer using the Add Fractions option. Four of the databases were downloaded in XML or flat text format from UniProt and indexed with various levels of complexity. The WHIM2 and WHIM16 sequence variant databases that were the same as used for the paper.

The High/High data were run using the five-step search shown in Figure 1, closely following the strategy reported in the supplementary information in Ntai et al. First, the High/High cRAWler performs Xtract deconvolution on the precursor and product ion spectra. The first search was performed against the mouse proteome with high proteoform complexity with narrow mass tolerances, identifying intact mouse proteoforms from the mouse host of the xenograft tissue. Any PrSMs with P-values lower than 1e-10 are passed to the next search through the Spectrum Confidence Filter node. The second search identified intact proteoforms of the sequence variants integrated with the rest of the information from the UniProt entries. While there are fewer protein entries in this database, the proteins in the database were indexed with high complexity. The third search identified intact proteoforms from a complex human proteome database. The fourth search detected truncated products of human proteins and the fifth search was an error tolerant delta-m search to find protein products that may significantly deviate in sequence or PTMs from the UniProt database.

Because the second step of the analysis requires a different database for the WHIM2 and WHIM16 samples, the processing workflow was cloned for the two different datasets. This is shown in Figure 2. The search results were subsequently combined into a single result using the consensus workflow.



Figure 2. Two processing workflows were used to search the WHIM2 and WHIM16 data with different databases in the second search as shown in Node 13 in Figure 1. The results of the two searches are integrated by the consensus step.



Low/High data analysis

The Low/High data were analyzed using a three-step search in a similar manner described for the High/High data. For these data, the workflow employed the Low/High cRAWler, which uses the ReSpect[™] deconvolution algorithm to determine average mass values for precursor proteins while still using Xtract deconvolution for the MS/MS spectra. In this search, there were three absolute mass nodes in sequence searching intact proteoforms for mouse, human, and human sequence variants. Since the MS1 data were acquired using ion trap mass spectrometry, a wider precursor tolerance had to be used (1000 ppm), but 10 ppm tolerance was still used for the high resolution MS/MS spectra.

Sliding window deconvolution and p-value calculations

All High/High and Low/High data were analyzed using sliding window deconvolution in a prerelease build of Biopharma Finder[™] 3.1 to produce mass, apex retention time, and abundance for the MS1 data. For the High/High data, the Xtract deconvolution algorithm was used to produce monoisotopic masses. The PrSMs with C scores >= 3 from the human searches were integrated with the sliding window deconvolution results using a Perl script. The integrated results were loaded into InfernoRDN (Pacific Northwest National Laboratories) and Microsoft Excel for LOESS normalization and calculation of p-values.

RESULTS

High/High WHIM2 versus WHIM16

The combined search results returned 30,104 PrSMs corresponding to 2,374 proteoforms and 602 protein isoforms. Table 1 below shows the list of PrSMs with a P-value for 1e-10 or better for each search shown in Figure 1.

Table 1. PrSM counts the WHIM2 and WHIM16 samples at each step in the ProSightPD workflow.

Search Order	Search type	# WHIM2 PrSMs	# WHIM16 PrSMs		
1	Mouse intact proteoforms	4063	3067		
2	Human sequence variants	89	53		
3	Human intact proteoforms	2611	2204		
4	Human truncations	48	236		
5	Human error tolerant	10709	7206		

There were significantly more PrSMs detected in the WHIM2 sample, indicating that there was more starting material for this sample than for WHIM16. Further, there were significantly more mouse proteoforms than human proteoforms, but this is likely the case because many of these mouse proteoforms have significant homology to those from human. Thus, many of these PrSMs may actually be human proteins, but it is not possible to distinguish whether or not they are mouse or human proteins. Finally, there are a lot of PrSMs identified in the final step, indicating that there are many proteoforms that differ from what is predicted from the UniProt entry.

The PrSM table was further filtered to those that are likely to be fully characterized by using a threshold C score of 3. Using the Proteome Discoverer software feature "Check All In This Table and All Associated Tables," all of the proteoforms and protein isoforms associated with these characterized PrSMs can be checked and filtered. There were 735 proteoforms with C scores 3 or more after filtering.

The "Found in Samples" columns were subsequently used to find proteoforms that were identified in only one of the two samples. Figure 3 shows the list of proteoforms to those that were only identified in the WHIM2 sample by filtering by "Not Found" in the WHIM16 sample. The list is sorted by decreasing numbers of PrSMs, which means the proteins at the top of the list with over 100 PrSMs are highly up-regulated in WHIM2 relative to WHIM16.

Figure 3. Proteome Discoverer view of proteoforms detected only in the WHIM2 sample.

	verer 2.3.0.267																	
View Administration	Tools Window Help																	
i) 🕼 🕼 🖡	; 🗊 👚 🥅 🔛	<u>ु</u>			Qi 🗧	×			N 🐗	2 🐻	4 🛃							
tart Page 🗙 Study: X	enograft 🗙 Study: Xenograft 🗙 Study 1_Hi	Hi_Absolute	Mass_WHIM2_WHIM16_Stier_	new X														
play Filter																		
Load 🛃 Save 💥 Clear	📕 Clear All 🥝 Apply 🧭 Cancel																	
N C Protein Isoform	Proteoforms																	
N - Proteoforms	Add group																	
N 📖 😑 PrSMs	Checked is true Remove																	
N MS/MS Spectr	Found in Samples has confi	dence Not I	Found (in sample) [S277] F280: S	ample Remove														
N Specialized Tr	1085																	
N - Annotation Pro	tein Groups																	
N 📖 🧿 Pathway Prote	in Groups																	
N C Result Statistic	s																	
Protein Isoforms Pr	oteoforms 😵 PrSMs 💱 MS/MS Spectrum	Info Inpu	ut Files 🧣 🛛 Specialized Trac	es Pathway P	rotein Groups	a 🧣 🛛 An	notation Pro	tein Gro	ups 🧣	Result Stati	stics							
E Checked Confi	dence Sequence	# PrSMs +	Theo. Mass [Da] Found in Samples	Confidence	by Search Eng	+ % Resid	Je Cleavages	(by 5 💽	Corrected D	Delta Mass (pp	m 💽 -Log P	Score (by S	earch Er 🛛	-Log E-V	alue (by !	Search E	r 📧 Fragment Map	
1 🗁 🔽 🤇	AGELTPEEEAQYKKAFSAVDTDGNGTINAQELC	109	15793.74165	High		31.25			0.53		81.61			73.69			View	
2 🗢 🗹	PNFSGNWKIIRSENFEELLKVLGVNVMLRKIAVA	104	15551.97990	High		17.65			-64.26		31.72			23.81			View	
3 🗢 🔽	PKRKAKGDAKGDKAKVKDEPQRRSARLSAKPA	69	9402.19256	High Med	um Low L	ow 29.55	9.09 4.55	4.55	0.79 0.	07 80,530.0	0.92 53.35	11.30 5	62 6.33	45.43	2.56	0.21 0	.00 View	
4 🗢 🔽 🤇	APPWVPAMGFTLAPSLGCFVGSRFVHGEGLRV	57	18654.56986	High		19.76			1.20		52.80			46.80			View	
5 😑 🗹 🤇	ADSELQLVEQRIRSFPDFPTPGVVFRDISPVLKD	39	19506.40230	High	Low	14.04	3.93		1.24	04,349.4	38.29	7	26	30.38		1.86	View	
6 🕂 🗹 🌔	AGELTPEEEAQYKKAFSAVDTDGNGTINAQELC	38	15751.73108	High		20.83			72,181.		25.15			19.75			View	
7 ቱ 🔽 🤇	AGNSILLAAVSILSACQQSYFALQVGKARLKYKV	31	16520.81025	High		13.10			0.56		29.55			21.63			View	
8 🗢 🔽 🤇	AKIKARDLRGKKKEELLKQLDDLKVELSQLRVAF	26	14453.53032	High	Low	15.70	7.44		0.31	2,906.1	29.75	4	67	21.83		0.00	View	
9 😔 🗹 🌔	MNLFRFLGDLSHLLAIILLLLKIWKSRSCAGISGK	26	24526.02868	High	Low	16.59	2.84		41.42	05,597.1	56.92	6	29	49.01		0.89	View	
10 🗢 🔽 🤇	MAGWNAYIDNLMADGTCQDAAJVGYKDSPSVA	25	15044.55603	High	Low	22.30	3.60		90.521.2	5,914.8	35.86	6	56	30.45		1.16	View	
11 🗢 🔽 🌔	PNFAGTWKMRSSENFDELLKALGVNAMLRKVA	22	15424.70746	High Med	um	14.81	7.41		-0.29 -1	.36	27.25	11.03		19.33	2.29		View	
12 🗢 🗹 🌔	MAGELTPEEEAQYKKAFSAVDTDGNGTINAQEL	21	15882.77157	High		13.10			17,293.3		25.11			19.71			View	
13 🗢 🔽 🌔	AAACRSVKGLVAVITGGASGLGLATAERLVGQC	19	26817.08257	High		13.13			36,306.	_	43.28		_	37.88			View	
14 🗢 🔽	SSGNAKIGHPAPNFKATAVMPDGQFKDISLSDY	17	22007.24892	High		19.80			0.44		51.95			44.04			View	
15 - 🗸	GGFFSSIFSSLFGTREMRILILGLDGAGKTTILYR	16	20315.40831	High	Medium	15.64	6.70	-	8.278.16	2,671.2	37.06	8	13	31.65		2.72	View	
17	ADQLIEEQVIEFKEAFSLFDKDGDGCITIRELG	15	10/90.92347	High	Madium	14.9/	10.2		0.85	E0 721	26.9/		12	19.06		2.71	View	
19	AAPOPPACEOGPEAPOPACOGOCCOPWACC	15	10022.01/01	High	meaium	13.66	10.26		-/5.23 507.27	03,721.	33.23	9	14	31.32	-	5.71	View	
19 - 7	PSIKLOSSOGEIEEV/DVEIAKOSV/TIVTHI EDLOL	10	18515 20427	High	-	16.77			53.93		20.73			20.32		-	View	
10 - V	ISEHI PINSRKCI REEIHKDI I VTGAYEISDOSGC	14	21736 10832	High	Low	6.95	4.81		49 919	50.483.9	25.79	6	52	20.33		1 12	View	
20	In the Instructure instruction (CISD0000C)		12007.01201	High	Low	11.72	5.47		80 839 1	-429.58	19.00	4	47	13.60	-	0.00	View	
20 -= 🔽	SGRGKQGGKARAKAKSRSSRAGI OFPVGRVHI	14	12077 (01201)						1-2,022.1	-LU.JO		- 4		20.55			View	
20 -= V 21 -= V 22 -= V	SGRGKQGGKARAKAKSRSSRAGLQFPVGRVHI ARTKQTARKSTGGKAPRKQLATKAARKSAPAT(14	15333.52110	High	LOW	15.67	0.47		-63.60		27.61							
20 +> V 21 +> V 22 +> V 23 +> V	SGRGKQGGKARAKAKSRSSRAGLQFPVGRVHI ARTKQTARKSTGGKAPRKQLATKAARKSAPAT(MNVDHEVNLLVEEIHRLGSKNADGKLSVKFGVL	14 14 14	15333.52110	High	Low	15.67	7.50		-63.60	0.59	27.61	5	69	18.97		0.00	View	
20 -> V 21 -> V 22 -> V 23 -> V 23 -> V 24 -> V	SGRGKQGGKARAKAKSRSSRAGLQFPVGRVHI ARTKQTARKSTGGKAPRKQLATKAARKSAPAT(MNVDHEVNLLVEEIHRLGSKNADGKLSVKFGVL APRKGKEKKEEQVISLGPQVAEGENVFGVCHIF	14 14 14 13	15333.52110 9092.87400 16131.49232	High High High	Low	15.67 30.00 9.40	7.50		-63.60 1.30 124.83	0.59	27.61 26.88 21.02	5.	69 57	18.97	-	0.00	View	
20 -20 -21 -22 -22 -22 -22 -22 -22 -22 -22 -22	SGRGKQGGKARAKAKSRSSRAGLQFPVGRVHI ARTKQTARKSTGGKAPRKQLATKAARKSAPAT(MNVDHEVNLLVEEIHRLGSKNAGGKLSVKFGVL APRKGKEKKEQVISLGPQVAEGEN/VFGVCHF MRYVASYLLAALGGKSSSAKDIKKILDSVGEA	14 14 14 13 13	15333.52110 9092.87400 16131.49232 11657.84731	High High High High	Low Low Medium	15.67 30.00 9.40 22.81	7.50 9.40 7.02		-63.60 1.30 124.83 -85.11	0.59	27.61 26.88 21.02 37.30	5.6	69 57 05	18.97 13.96 29.39		0.00	View View View	
20	SGRGKQGGKARAKAKSRSSRAGLQFPVGRVHI ARTKQTARKSTGGKAPTKQLATKAARKSAPAT(MMVDHEVNLLVEEHRLGSNNJDGKLSVKFGVL APRKGEKKEQVISLGPQVAEGENVFGVCHIF MRYVASYLLAALGGNSSPSACDIKKLDSVGIEA AATMKKAAAEDVIVTFEDQKINKFARTISTRIT	14 14 14 13 13 12	15897.81361 15333.52110 9092.87400 16131.49232 11657.84731 15215.54998	High High High High High	Low Low Medium Low	15.67 30.00 9.40 22.81 10.61	7.50 9.40 7.02 3.79		-63.60 1.30 124.83 -85.11 -65.78	0.59 8.838.4 36.754.4 33.890.5	27.61 26.88 21.02 37.30 22.06	5 6 9	69 57 05 34	18.97 13.96 29.39 14.15		0.00 1.16 3.65 0.94	View View View View View	
20	SGRGKQGGKARAKAKSRSSRAGLQFPVGRVHI ARTKQTARKSTGGKAPRKQLATKAARKSAPAT(MIVDHEVULVEIHHLGSKADAGKSVFGVU- APRKGKEKKEEQVISLGPQVAEGENVFGVCHIF MRYVASYLLAALGGKGSSPAXKDIKKILDSVIEA AATMKGAABDWIVTFEDQQKINKFARNTSRIT PEPSKSAPAPKKGSKKATKAQKKOCKKRKRSF	14 14 13 13 12 12	15373, 3138 15333, 52110 9092, 87400 16131, 49232 11527, 54998 13810, 54322	High High High High High High	Low Low Medium Low Medium L	15.67 30.00 9.40 22.81 10.61 pw 15.32	7.50 9.40 7.02 3.79 7.26	7.26	-63.60 1.30 124.83 -85.11 -65.78 58.396.1	0.59 8.838.4 36.754.4 33,890.5 78.787.1	27.61 26.88 21.02 37.30 22.06 4.12 23.37	5 6 9 6 10	69 57 05 34 .15 4.54	18.97 13.96 29.39 14.15 17.97		0.00 1.16 3.65 0.94 4.75 C	View View View View View View View	

In total, there were 78 proteoforms with five or more PrSMs that were identified only in the WHIM2 sample, while there were 75 proteoforms with five or more PrSMs that were identified only in the WHIM16 sample. Table 2 shows the top five proteoforms by PrSM count unique to either the WHIM2 or WHIM16 sample. While many of these proteoforms are for protein isoforms only identified in one of the two samples, Figure 4 shows example of a protein that was identified in both samples. In this case, there were two proteoforms only identified in both samples and two proteoforms identified only in the WHIM2 sample. The most abundant proteoform was unmodified while the others were acetylated at three different sites.

Table 2. Top five by PrSM unique proteoforms for each sample.

Sample	Proteoform Mass	Best C score	Protein entry	# PrSMs	Best log E-value
WHIM2	15793.74	1514	Calmodulin-like protein 5	109	73.69
WHIM2	15551.98	814	Cellular retinoic acid-binding protein 2	104	23.81
WHIM2	9402.193	697	High mobility group nucleosome- binding domain-containing protein 4	69	45.43
WHIM2	18654.57	1701	NP_000705-variant:Translocator protein	57	46.80
WHIM2	19506.40	583	Adenine phosphoribosyltransferase	39	30.38
WHIM16	10304.19	1063	Protein S-100P	60	22.3
WHIM16	12635.34	821	SH3 domain-binding glutamic acid-rich-like protein	59	51.27
WHIM16	9546.343	1537	Triple QxxK/R motif-containing protein	48	32.79
WHIM16	10393.22	571	Protein S-100P	46	28.46
WHIM16	7760.944	740	Adipogenesis regulatory factor	43	82.17

Figure 4. 60S ribosomal protein L35 shows two proteoforms identified in both samples and two proteoforms identified only in the WHIM2 sample. Three of these proteoforms have the same mass, differing only by the site of acetylation.



Low/High WHIM2 versus WHIM16

The search of the Low/High data produced more than 17000 PrSMs for 48 protein isoforms and 254 proteoforms. 654 PrSMs had C scores greater than 3, corresponding to 38 proteoforms and 24 protein isoforms. Of the 38 confidently-characterized proteoforms, 15 were larger than 40 kDa demonstrating the power of the low/high method for identification of larger proteins. Many of those 24 confidently characterized protein isoforms were for mouse proteins. However, there were some human proteins uniquely identified in each sample. Transmembrane protein 43 (44 kDa, 20 PrSMs), high mobility group nucleosome-binding domain-containing protein (31 kDa, 5 PrSMs), and reticulon-4 (20.4 kDa, 442 PrSMs) were unique to the WHIM2 sample and parathymosin (11.5 kDa, 2 PrSMs) was unique to the WHIM16 sample.

It is likely that the requirement for wider precursor tolerances reduced the sensitivity of the approach. The most abundant protein in both samples, glyceraldehyde 3-phosphate dehydrogenase, had over 5000 PrSMs between the two samples. However, none of the PrSMs were identified with a non-zero C score. The 1000 ppm precursor mass tolerance is equivalent to +/- 36 Da, which is too wide to distinguish between many of the possible combinations of PTMs. These data would benefit from a "medium/high" acquisition strategy where the precursor spectrum was acquired at 15K resolution. In this case, we could use a much narrower precursor mass tolerance (~30-50 ppm), decreasing the number of theoretical proteoforms to be compared in the database. Further, the majority of the fragment coverage from HCD of this protein was near the N- and C-termini. Another MS/MS mode such as CID will likely produce more fragmentation in the middle of the protein, which would lead to improved C scores.

Sliding window deconvolution of High/High data

Sliding window deconvolution of the WHIM2 and WHIM16 data shows that we should expect significant differences between even common proteoforms between the samples. Figure 5 shows a mirror plot of the sliding window deconvolution results for the first GELFrEE fractions of the WHIM2 (upper plot) and WHIM16 (lower plot) samples.

Figure 5. Mirror plot comparison of GELFrEE fraction 1 for the WHIM2 and WHIM16 samples.



All 18 High/High datasets were analyzed using sliding window deconvolution and integrated with the PrSMs as previously explained. The WHIM2 and WHIM16 data had three replicates each and thus p-values could be calculated. Of the 735 confidently characterized proteofroms, more than 400 proteoforms exhibited a significant difference in expression with p-values <0.01. This is over half of the confidently identified proteoforms, which is higher than one would normally see by quantitative bottom up proteomics.

CONCLUSIONS

- ProSightPD is a powerful tool for analysis of complex top down proteomics data.
- The C-score is an especially important measurement for proteoform identification.
- Integration of sliding window deconvolution results showed the promise for quantitative label-free top down protein ID.

REFERENCES

- 1. Ntai, I et al, Mol Cell Proteomics, 2016, 15, pp 45-56.
- 2. LeDuc, R. D., Fellers, R. T., Early, B. P., Greer, J. B., Thomas, P. M., Kelleher, N.L., J. Proteome *Res.* **2014**, *13*, 3231-3240.
- 3. Fellers, R. T., Greer, J. B., Early, B. P., Yu, X., LeDuc, R. D., Kelleher, N. L. Proteomics, 2015, 15, pp. 1235-1238.

ACKNOWLEDGEMENTS

Thank you to Torsten Ueckert and Carmen Paschke for their continued assistance with the development of the ProSightPD nodes.

TRADEMARKS/LICENSING

© 2018 Thermo Fisher Scientific Inc. All rights reserved. ProSightPC and ProSightPD are trademarks of Proteinaceous, Inc. Microsoft is a registered trademark of Microsoft Corporation. ReSpect is a trademark of Positive Probability, Ltd. All other trademarks are the property of Thermo Fisher Scientific and its subsidiaries. This information is not intended to encourage use of these products in any manner that might infringe the intellectual property rights of others. **Thermo Fisher**



WHIM2

16619.908 17910.8520762.826 16619.904

20000

WHIM16



PO65237-EN 0518S