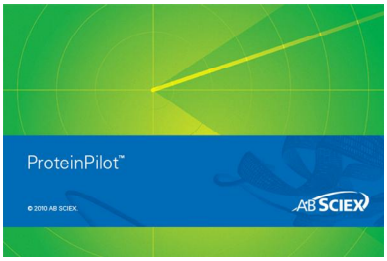
ProteinPilot™ Descriptive Statistics Template

Powerful Analysis Tools for Protein Identification Results from ProteinPilot™ Software

Sean L Seymour, Christie Hunter
AB SCIEX, USA

Powerful mass spectrometers like the TripleTOF™ 5600 System can rapidly generate extremely large amounts of data in short order. For today's researchers, tools that can logically and efficiently distill the massive amounts of data down into easily interpretable results are critical. ProteinPilot™ Software is a powerful, robust, easy to use software tool for protein identification and quantification for discovery research and protein characterization. With its hybrid sequence tag and database search approach using feature probabilities, the powerful Paragon™ Algorithm can search for hundreds of modifications and sequence variants in a single search[1]. Coupled with the Pro Group™ Algorithm for protein inference analysis, peptide results are condensed down to the most defensible set of detected proteins with ambiguity among multiple accession numbers reported when appropriate. Finally, an integrated false discovery rate (FDR) analysis gives a rigorous report, detailing the quality of protein and peptide identifications[2].

The understanding of experimental results involving peptide and protein identifications needs more than just a simple list of proteins, however long and impressive it may be. There are many different types of post-acquisition analysis that can be performed that are highly valuable to the protein researcher to ensure results quality and enable workflow refinement. Many of these types of analysis have been combined into a single Excel-based processing tool, the ProteinPilot™ Descriptive Statistics Template (PDST). The PDST tool automatically generates a wealth of important information from data-intensive proteomics experiments, which would normally require many weeks of manual data crunching.



Key Features of the ProteinPilot™ Descriptive Statistics Template

- Simply cut-and paste false discovery rate analysis results and protein and peptide exports from Paragon™ Algorithm searches within ProteinPilot™ Software into the PSDT excel spreadsheet, and Click 'Calculate' to generate the full analysis.
- Enables the rapid assessment of the quality of identification and quantification.
- Enables the characterization of sample preparation – digestion quality, modification frequencies, labeling efficiency, etc.
- Enables the optimization of acquisition parameters using detailed metrics on acquisition redundancy, chromatography, mass accuracy, etc.
- Generate volcano plots and compute false discovery rates of differential expression for simple quantitation studies.
- Virtually all quantitative metrics (>7000 data points) are captured in a single column that can be saved for future use – from simple comparison to complex data mining.

Protein N	Representative Accession	Species	Name	Unused ProtScore	Total ProtScore	Ambiguous Accessions	Confident Peptides	Confident Sequences	Confident Spectra	%Seq Cov - peptides any conf	%Seq Cov - peptides >50% conf	%Seq Cov - peptides ≥95% conf	Peptides(95%)
1	P36683|ACON2_ECOLI	ECOLI	Aconitate hydratase 2 - Escherichia coli (strain K12)	68.17	68.17	1	63	37	248	52.2	43.5	41.4	58
2	P0A8V2|RPOB_ECOLI	ECOLI	DNA-directed RNA polymerase subunit beta - Escherichia coli (strain K12)	61.45	61.45	1	37	34	70	47.1	28.3	23.8	34
3	P0A6N1|EFTU_ECOLI	ECOLI	Elongation factor Tu - Escherichia coli (strain K12)	61.18	62.52	1	95	36	739	86.3	73.9	70.8	95
4	P0A8T7|RPOC_ECOLI	ECOLI	DNA-directed RNA polymerase subunit beta' - Escherichia coli (strain K12)	60.19	60.33	1	47	38	94	44.4	29.5	24.7	43
5	P0A853|TNAA_ECOLI	ECOLI	Tryptophanase - Escherichia coli (strain K12)	54.34	54.34	1	55	28	259	63.1	51.0	47.4	49
6	P0A6Y8|DNAK_ECOLI	ECOLI	Chaperone protein dnaK - Escherichia coli (strain K12)	51.62	51.77	1	42	26	105	53.8	45.3	40.9	42
7	P0A836|SUCC_ECOLI	ECOLI	Succinyl-CoA ligase [ADP-forming] subunit beta - Escherichia coli (strain K12)	50.09	50.72	1	45	27	229	77.1	63.1	58.2	39
8	P02787|TRFE_HUMAN	HUMAN	Serotransferrin precursor - Homo sapiens	48.97	48.97	1	37	25	103	54.7	39.1	35.0	35
9	P02769|ALBU_BOVIN	BOVIN	Serum albumin precursor - Bos taurus	48.57	48.57	1	35	25	115	56.7	44.2	41.0	35
10	P0A705|IF2_ECOLI	ECOLI	Translation initiation factor IF-2 - Escherichia coli (strain K12)	48.33	50.82	1	31	27	47	52.1	35.5	29.0	28
11	P0A6F5|CH60_ECOLI	ECOLI	60 kDa chaperonin - Escherichia coli (strain K12)	47.33	47.91	1	52	25	302	63.7	50.2	45.6	49
12	P00761|TRYP_PIG	PIG	Trypsin precursor - Sus scrofa	47.16	47.86	1	157	38	1243	62.8	55.0	55.0	176
13	P0A9B2|G3P1_ECOLI	ECOLI	Glyceraldehyde-3-phosphate dehydrogenase A - Escherichia coli (strain K12)	46.66	47.10	1	54	28	235	66.2	57.4	57.4	53
14	P00722|BGAL_ECOLI	ECOLI	Beta-galactosidase - Escherichia coli (strain K12)	46.43	46.72	1	36	25	83	35.4	25.6	23.1	34

Figure 1. Reporting Summary Tables for Protein and Peptide Identifications. Protein and peptide tables are provided to cleanly organize identification and quantitation results. Several additional metrics not computed in ProteinPilot™ Software are provided.

Usage and Basic Reporting

To use the PDST tool, three data elements that are generated from a ProteinPilot™ Software search must be pasted into the Excel template – the FDR analysis, the protein export, and the peptide export. Once this information is in place, the entire workbook is calculated and multiple analyses are automatically performed and organized into different worksheets. First, protein and peptide summary tables are created that distill all the information down into a two dimensional table, providing informative formatting and additional information not included in the exports (Figure 1).

Acquisition Characterization

One of the keys to fully optimizing the quality of data acquired by a mass spectrometer is the ability to measure the appropriate quantitative metrics. The PDST provides this from many angles: basic descriptors like the distributions of charge state, mass, and m/z for confidently identified peptides (Figure 2, top), analysis of the redundancy of acquisition (repeated selection and fragmentation of the same physical peptide), and metrics breakdown by fraction for 2D workflows. With ProteinPilot™ Software 4.0, the peptide intensity at the LC peak apex is reported, which enables several additional and highly valuable analyses to be performed. For example, the distribution of precursor signals is computed, which directly measures the dynamic range of peptides identified in the acquired data (Figure 2, bottom).

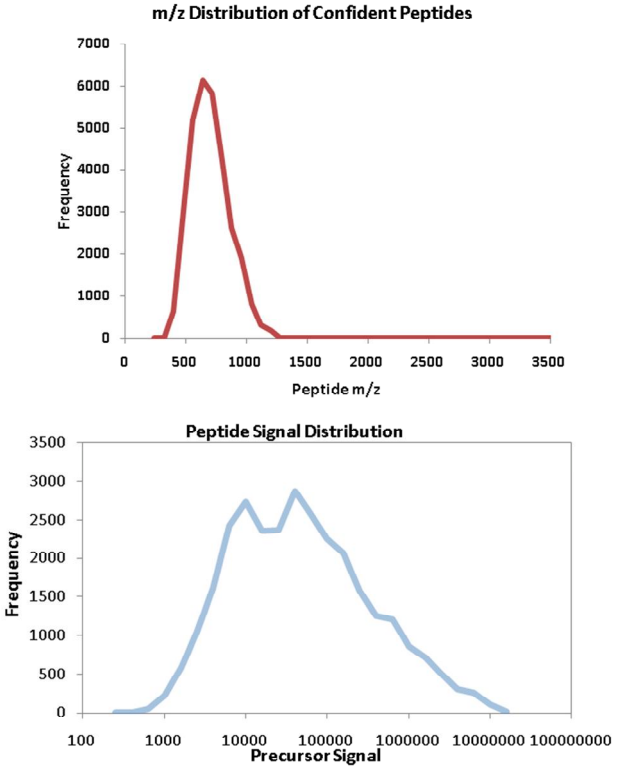
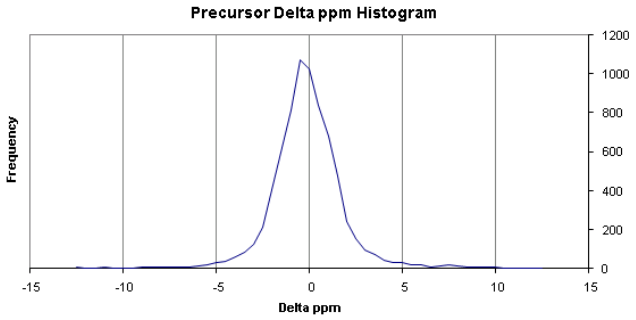


Figure 2. Acquisition Characterization. Descriptive analyses are done considering only peptides identified at <5% local FDR. (Top) Precursor mass/charge distribution intensities of peptide precursors observed in a complex cell lysate using the TripleTOF™ 5600 system. (Bottom) MS intensity distribution peptide precursors showing the dynamic range of peptides identified in a human plasma sample, almost 5 orders of dynamic range.



Mass Error Summary Statistics Table

	Std. Deviation	RMS	Average Error
Delta m/z error	0.00081	0.00082	-0.00011
Delta ppm error	1.26	1.27	-0.19
Delta Sqrt m/z error	1.53E-05	1.54E-05	-2.23E-06

Figure 3. Precursor Mass Accuracy Evaluation. The top figure shows a histogram of the scatter of ppm error, while the table at the bottom reports metrics on precision, bias, and the combined RMS metric. The PDST also analyzes mass accuracy in many other ways, including as a function of retention time and as a function of peptide intensity.

Analysis of Mass Accuracy

ProteinPilot™ Software has always used an automatic recalibration of both MS and MS/MS data independently by doing a rapid test search, getting a significant number of confident identifications, followed by the use of these to remove any observed mass shift in the data. In ProteinPilot™ Software 4.0, the precursor mass is determined by looking across the full elution of the peptide, rather than just considering the closest survey spectrum. This gives a demonstrable improvement in mass accuracy over previous versions. The PDST also provides a detailed graphical and numerical assessment of precursor mass accuracy, allowing easy assessment of the mass accuracy in the data (Figure 3). Analysis as a function of retention time allows assessment of the time stability of an instrument's calibration, while analysis as a function of peptide intensity gives a description of mass accuracy that allows comparison even across samples with different complexity.

Identification Feature Characterization

Proteases do not have perfect cleavage specificity. Thus, the ability of a search engine to search for missed cleavages (under cleavage) and unexpected cleavages (over cleavage) is extremely important to the fidelity of identification results. The unique features of the Paragon Algorithm enable searching both these types of digestion deviations, in addition to hundreds of sample preparation and biological modifications. As we dig deeper into our proteomic samples with faster and more sensitive instruments, more of the low level features will be detected and must be considered in our searches. Failure to consider relevant features can cause false positive hits to other peptides, causing false positive proteins as well.

The PDST enables the detailed characterization of both digestion and modification features. Monitoring the missed cleavage and semi-tryptic rates observed in each study is an effective way to ensure that the digestions are working well and reproducibly. Unusual missed cleavage or semi-tryptic rates may signal a problem (Figure 5, top). A highly detailed view of digestion is also provided. The heat map (Figure 5) shows the cleavage rates observed between each residue pair.

The PDST also provides a list of the 25 most frequent modifications (including substitutions), giving the number of occurrences in confidently identified peptides. It also computes the fraction of total ion signal having the modification of all forms of the same base sequences, as measured via peptide elution apex intensities (Figure 4, bottom). This allows for the rigorous QC of sample preparation steps, like cysteine alkylation, and labeling chemistries. This can also be very useful for tracking the rates of undesirable artifact modifications such as oxidation, carbamylation, and the sometimes inevitable side reactions that occur in sample preparation steps.

Support for All MS Instruments

ProteinPilot™ Software can process data from all other MS instruments, as long as the data is converted into mgf format (mascot generic format). Once the data is processed, protein and peptide exports can be created and pasted into the PDST. Most of the qualitative plots are automatically created as discussed here, for detailed analysis and comparison of other MS instrument data.

Missed Cleavages (Under-cleavage)

Missed Cleavages	*Spectra in Selected Set*	*% of Selected Set*
0	13129	95.5%
1	623	4.5%
2	1	0.0%
3	1	0.0%
4	0	0.0%
5	0	0.0%

Expected Digestion (Over-cleavage)

Peptide Type	*Spectra in Selected Set*	*% of Selected Set*
Expected termini	13053	94.9%
Semi-specific (only one expected terminus)	573	4.2%
Non-specific (neither terminus expected)	128	0.9%

Most Frequent Single Features

Rank	Feature	Exact Delta	#	Fraction of Sequence Signal
1	iTRAQ8plex@N-term	304.2054	19168	0.961
2	iTRAQ8plex(K)	304.2054	14006	0.994
3	Oxidation(M)	15.9949	4306	0.684
4	Deamidated(N)	0.9840	1687	0.189
5	Deamidated(Q)	0.9840	1171	0.163
6	Methylthio(C)	45.9877	781	0.972
7	Cation:K(E)	37.9559	382	0.030
8	iTRAQ8plex(Y)	304.2054	258	0.030
9	iTRAQ8plex(S)	304.2054	207	0.012
10	Gln->pyro-Glu@N-term	-17.0265	155	0.194
11	Cation:K(D)	37.9559	123	0.019
12	Methyl(H)	14.0157	121	0.008
13	Delta:H(2)C(2)@N-term	26.0157	98	0.004
14	Oxidation(P)	15.9949	42	0.005
15	Ammonia-loss(N)	-17.0265	26	0.006
16	iTRAQ8plex(T)	304.2054	25	0.003
17	Acetyl@N-term	42.0106	17	0.000
18	Methyl(E)	14.0157	17	0.001
19	Dioxidation(M)	31.9898	16	0.004
20	Oxidation(W)	15.9949	15	0.007
21	Deamidated(R)	0.9840	14	0.003
22	Methyl(D)	14.0157	9	0.002
23	Oxidation(D)	15.9949	9	0.001
24	Amino(Y)	15.0109	7	0.003
25	Oxidation(H)	15.9949	7	0.001

Figure 4. Characterization of Features Identified in Search. (Top and middle) Digestion frequencies are useful to monitor as deviations from normal can indicate problems with sample preparation. (Bottom) The most frequent modification table is also useful for sample QC. For example, here we can see that methylthio modification from MMTS alkylation labeled Cys 97.2% as measured by peptide signal.

Conclusions

- The ProteinPilot™ Descriptive Statistics Template is a powerful tool to provide a much deeper understanding of MS identification and quantitation results, in minutes rather than weeks.
- Many detailed views are provided by the PDST that help characterize the quality of collected data on the mass spectrometer, including mass accuracy analysis, acquisition redundancy analysis, mass and charge state distributions, etc.
- The characterization of modification and digestion frequencies provided by the PDST can be a great quality control tool.
- Basic support for analyzing label-based quantitative experiments is provided, including metrics and graphical views of variation, volcano plots, and calculation of the false discovery rate of differential expression for some workflows.
- >7000 quantitative readout are produced by the PDST, which can be captured in a single column, enabling everything from simple comparison of two data sets to complex data mining.

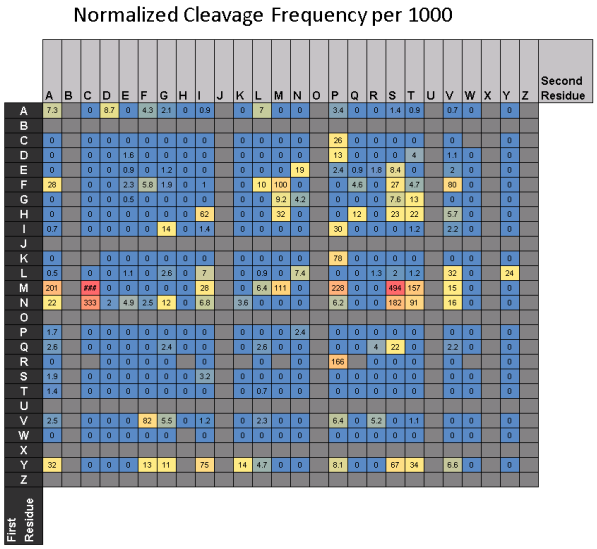


Figure 5. Digestion Frequencies. Shown are the observed frequencies of cleavages for a sample digested with trypsin. The frequency for each residue pair is computed as the number of observed cleavages divided by the number of possible sites (reported as frequency per 1000).

References

1. The Paragon Algorithm, a Next Generation Search Engine That Uses Sequence Temperature Values and Feature Probabilities to Identify Peptides from Tandem Mass Spectra, Shilov IV *et al.* (2007), *Mol. Cell. Proteomics,* **6**, 1638-1655.
2. Nonlinear Fitting Method for Determining Local False Discovery Rates from Decoy Database Searches. Tang W et al. (2008), J. Prot. Res. 7(9), 3661–3667.
3. The ProteinPilot™ Descriptive Statistics Template can be downloaded from the website - http://www.absciex.com/PDST

Publication number: 1910211-02

Headquarters
110 Marsh Drive | Foster City CA 94404 USA
Phone 650-627-2600
www.absciex.com

International Sales
For our office locations please call the division headquarters or refer to our website at
www.absciex.com/offices