Supplementary Information for

Improved Accuracy for Label-free Absolute Quantification of Proteome by Combining Absolute Protein Expression Profiling Algorithm and Summed Tandem Mass Spectrometric Total Ion Current

Qi Wu, Yichu Shan, Yanyan Qu, Hao Jiang, Huiming Yuan, Jianxi Liu, Shen Zhang, Zhen Liang, Lihua Zhang,* and Yukui Zhang

1. Experimental section for UPS2 dataset

1.1 Materials and Reagents

UPS2 Proteomics Dynamic Range Standard, urea (99.5%), formic acid (FA) and protease inhibitor complex cocktail were purchased from Sigma-Aldrich (St. Louis, MO). Modified trypsin, sequencing grade, was ordered from Promega (Madison, WI). Dithiothreitol (DTT) and iodoacetamide (IAA) were bought from Acros (Morris Plains, NJ). Acetonitrile (ACN, HPLC grade) was obtained from Merck (Darmstadt, Germany). Water was purified by a Milli-Q system (Millipore, Milford, MA).

Fused-silica capillaries (75 and 150 μm i.d. × 375 μm o.d.) were purchased from Sino Sumtech (Handan, China). BCA protein assay kit was ordered from Beyotime Institute of Biotechnology (Tianjin, China). Venusil XBP C18 and C8 particles (5 μm, 120 Å) were bought from Bonna-Agela Technologies (Tianjin, China). ReproSil-Pur C18-AQ particles (5 μm, 120 Å) were obtained from Dr. Maisch GmbH (Ammerbuch-Entringen, Germany).

1.2 Sample Preparation

Rat brain tissue was cut into small pieces, then washed with cold PBS three times, and finally grounded to powder in liquid nitrogen. Yeast cells were harvested and washed with cold PBS three times as well. Then they were respectively suspended in the extraction buffer composed of 8 M urea and 1% (v/v) protease inhibitor cocktail. Rat brain powder suspension was first homogenized for 60 s at 10000 rpm, and then ultrasonicated for 100 s at 130 w in ice-bath, while yeast cell

suspension was directly ultrasonicated for 300 s. They were finally centrifuged at 25 000 rpm for 40 min. The resulting supernatants were desalted with a home-packed C8 trap column (XBP) and lyophilized by a Refrigerated CentriVap Concentrator (LABCONCO, Kansas City, MO). After resuspending them in 50mM NH_4HCO_3, the protein concentrations were determined by BCA assay.

1.3 Protein Digestion

The UPS2 standard, yeast and rat protein extracts were respectively denatured by heating to 90℃ for 20 min, reduced by 10 mM DTT at 56 °C for 2 h, and alkylated by 25 mM IAA in the dark at room temperature for 40 min. Then trypsin was added with a substrate-to-enzyme ratio of 25:1 for yeast and rat, and 50:1 for UPS2 standard, followed by incubation at 37 °C for 16 h. Then 1% (v/v) FA was added to terminate the digestion. Finally, the peptide solutions were further centrifuged at 20 000 rpm for 20 min and stored at -80 °C until MS analysis.

1.4 nano-RPLC-ESI-MS/MS

The analyses for protein digests of UPS2 dataset were performed on both LTQ XL and Orbitrap Velos mass spectrometers (Thermo Fisher, San Jose, CA, USA). Besides the mass spectrometer in use, the nano-RPLC-ESI-MS/MS system was composed of Accela 600 pump and Accela Autosampler (Thermo Fisher, San Jose, CA, USA). The sample loading amount for yeast and rat digests was 1 μg. For UPS2, the sample was diluted 100 times to give a final dynamic range of 500 fmol - 5 amol on column. Samples were loaded onto a 3 cm long homemade trap column(XBP C18, 150 μm i.d.) and eluted onto a 15 cm long home packed separation column (AQ C18, 75 μm i.d.) using a 90 min gradient under the flow rate of 200 nL/min.

The mass spectrometers were operated at positive ion mode. The temperature of ion transfer capillary was set at 200 ℃, the spray voltage was set at 2.1 kV and the normalized collision energy was set at 35%. Total ion chromatograms were recorded from m/z 300 to 1800. MS/MS spectra were acquired in the data dependent mode in which the 7 (LTQ) or 15 (Orbitrap) strongest species were selected for fragmentation by CID. MS/MS scans consisted of 1 micro scan with an AGC target of 10000. The dynamic exclusion settings were as follows: repeat count: 2 for LTQ and 1 for Orbitrap; repeat duration: 60 s for LTQ and 40 s for Orbitrap; exclusion list size: 50 for LTQ and 500 for Orbitrap; exclusion duration: 180 s for LTQ and 40 s for Orbitrap.

2. Comparison of Random Forest, RIDOR and J4.8 Decision Trees

The estimators to evaluate the performance of three machine learning algorithms (Random Forest, RIDOR, J4.8 Decision Trees) using five different arff files are summarized in Table S1. All of them are closer to 1 the better. J4.8 Decision Trees are behind Random Forest in every angle, while RIDOR surmounted Random Forest in recall of observed peptides. However, the other two estimators of RIDOR are significantly inferior to Random Forest, especially the percentage of correctly classified instances, deteriorating its overall performance. Therefore, Random Forest is our final choice for training.

Table S1. Comparison of Random Forest, RIDOR and J4.8 Decision Trees

Algorithm	Random Forest weka.classifiers.meta.CostSensitiveClassifier -cost-matrix "[cost matrix]" -S 1 -W weka.classifiers.meta.Bagging -- -P 100 -S 1 -I 10 -W weka.classifiers.trees.RandomForest -- -I 10 -K 5 -S 1					RIDOR weka.classifiers.meta.CostSensitiveClassifier -cost-matrix "[cost matrix]" -S 1 -W weka.classifiers.meta.Bagging -- -P 100 -S 1 -I 10 -W weka.classifiers.rules.Ridor -- -F 3 -S 1 -N 2.0					J4.8 Decision Trees weka.classifiers.meta.CostSensitiveClassifier -cost-matrix "[cost matrix]" -S 1 -W weka.classifiers.meta.Bagging -- -P 100 -S 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.25 -B -M 2				
Source of arff file	LTQ-yeast	LTQ-rat	Orbitrap-yeast	Orbitrap-rat	Orbitrap@86-yeast	LTQ-yeast	LTQ-rat	Orbitrap-yeast	Orbitrap-rat	Orbitrap@86-yeast	LTQ-yeast	LTQ-rat	Orbitrap-yeast	Orbitrap-rat	Orbitrap@86-yeast
Percentage of correctly classified instances	92.3%	94.3%	91.5%	92.4%	95.7%	87.5%	91.7%	85.0%	87.3%	94.3%	91.2%	94.1%	90.8%	92.0%	95.3%
F-measure of observed peptides	0.713	0.672	0.691	0.654	0.712	0.635	0.615	0.598	0.569	0.668	0.685	0.670	0.668	0.639	0.693
Recall of observed peptides	0.758	0.748	0.731	0.731	0.840	0.859	0.850	0.862	0.853	0.900	0.751	0.770	0.718	0.722	0.828

3. Relative deviation of quantification for every UPS2 protein for yeast-training and rat-training data

The relative deviation of quantification for every UPS2 protein for yeast-training and rat-training data on both LTQ XL and Orbitrap Velos are summarized in Table S2 and S3. The averages of their absolute values were calculated.

Table S2. Relative deviation of quantification for every UPS2 protein for yeast-training and rat-training data on LTQ XL

Protein accession number	Injected amount (fmol)	Trained by yeast			Trained by rat		
		APEX	APEX-SI	APEX-SMT	APEX	APEX-SI	APEX-SMT
P00709ups	5	314%	291%	-25%	390%	372%	-14%
P02753ups	5	14%	36%	-80%	13%	38%	-80%
P06732ups	5	134%	218%	-57%	154%	249%	-55%
P12081ups	5	121%	180%	-55%	123%	188%	-56%
P16083ups	5	208%	259%	-55%	170%	216%	-62%
P61626ups	5	139%	191%	-41%	145%	205%	-41%
P00167ups	50	172%	605%	62%	172%	619%	58%
P01133ups	50	264%	47%	29%	310%	71%	41%
P02144ups	50	118%	195%	-36%	121%	205%	-37%
P04040ups	50	93%	213%	-13%	77%	192%	-22%
P15559ups	50	39%	99%	-46%	51%	121%	-43%
P62937ups	50	202%	186%	-6%	113%	106%	-35%
P63165ups	50	33%	78%	-57%	32%	79%	-58%
Q06830ups	50	174%	257%	2%	167%	258%	-2%
P00915ups	500	13%	-31%	25%	21%	-25%	30%
P00918ups	500	-34%	-16%	-26%	-22%	1%	-14%
P01031ups	500	-4%	-49%	40%	-13%	-53%	23%
P02768ups	500	20%	-16%	159%	27%	-9%	166%
P41159ups	500	3%	62%	-32%	-27%	18%	-53%
P62988ups	500	-37%	-45%	-34%	-16%	-26%	-15%
P68871ups	500	-35%	-51%	-66%	-40%	-54%	-69%
P69905ups	500	-45%	-33%	-54%	-43%	-29%	-53%
Average of absolute values		***101%***	***144%***	***45%***	***102%***	***143%***	***47%***

Table S3. Relative deviation of quantification for every UPS2 protein for yeast-training and rat-training data on Orbitrap Velos

Protein accession number	Injected amount (fmol)	Trained by yeast			Trained by rat		
		APEX	APEX-SI	APEX-SMT	APEX	APEX-SI	APEX-SMT
O76070ups	0.5	7252%	3684%	629%	7554%	3906%	636%
P01008ups	0.5	664%	1166%	101%	671%	1200%	96%
P01344ups	0.5	4990%	389%	413%	5378%	439%	436%
P08263ups	0.5	4712%	524%	230%	4439%	502%	201%
P55957ups	0.5	2166%	460%	97%	2015%	433%	78%
P61769ups	0.5	5035%	10760%	1488%	4516%	9863%	1285%
P00709ups	5	1245%	625%	102%	1231%	632%	93%
P02753ups	5	1284%	942%	188%	1189%	894%	159%
P06732ups	5	914%	193%	-4%	868%	185%	-11%
P12081ups	5	975%	294%	16%	1026%	320%	18%
P16083ups	5	419%	60%	-57%	368%	47%	-62%
P61626ups	5	578%	74%	-27%	710%	112%	-15%
P63279ups	5	1327%	415%	24%	1240%	391%	12%
Q15843ups	5	659%	391%	-14%	676%	415%	-14%
P00167ups	50	351%	1104%	354%	308%	1014%	298%
P01133ups	50	81%	154%	-61%	77%	153%	-63%
P02144ups	50	120%	520%	-4%	116%	520%	-8%
P04040ups	50	124%	166%	-9%	109%	152%	-18%
P15559ups	50	95%	326%	-6%	81%	302%	-16%
P62937ups	50	291%	85%	-3%	198%	44%	-28%
P63165ups	50	123%	436%	20%	108%	412%	9%
Q06830ups	50	326%	454%	61%	323%	462%	55%
P00915ups	500	-6%	-39%	6%	-16%	-44%	-9%
P00918ups	500	-34%	-45%	-8%	-29%	-40%	-4%
P01031ups	500	-36%	-63%	-6%	-40%	-65%	-15%
P02768ups	500	13%	7%	171%	18%	13%	174%
P41159ups	500	-54%	-67%	-75%	-59%	-70%	-79%
P62988ups	500	-9%	-52%	37%	17%	-38%	70%
P68871ups	500	-55%	-77%	-81%	-50%	-73%	-79%
P69905ups	500	-72%	-39%	-85%	-72%	-37%	-85%
Average of absolute values		***1134%***	***787%***	***146%***	***1117%***	***759%***	***138%***

4. Parameters and diagnostic plots of power law global error model (PLGEM)

Two important parameters were set as follows: trimAllZeroRows=TRUE and zeroMeanOrSD="trim" to reduce the side effect of missing values normally seen in proteomics dataset. The number of iterations of the permutation step was set to 2000 rather than the default 500 to stabilize p values from run to run. All other parameters were default.

The PLGEM diagnostic plots with variance-versus-mean trend line for the three algorithms are summarized in Figure S1, from which we could see that APEX-SI and APEX-SMT fitted to the model very well.

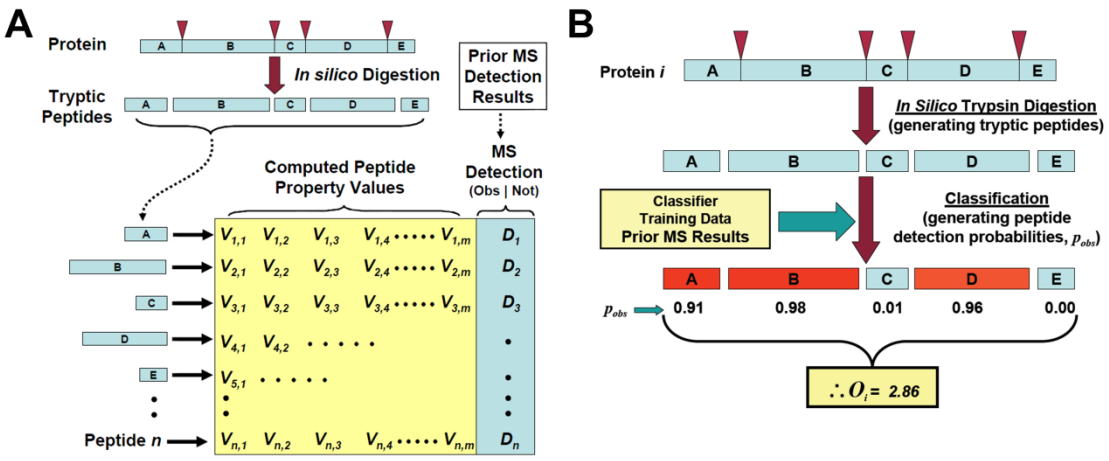

Figure S1. The processes of training data construction (A) and Oi value calculation (B). (A) Protein sequences from database for training data search undergo an in-silico trypsin digestion to form a set of peptides. A number (m) of peptide physicochemical properties are calculated for each peptide (for peptide i, properties *1-m* are denoted in the figure as: $V_{i,1}$, $V_{i,2}$, $V_{i,3}$,... $V_{i,m}$) and prior identification results of training data are searched to determine if the peptide has been observed or not (for peptide i, the detection call is denoted in the figure as D_i). This matrix forms the arff file. (B) Each protein undergoes an in-silico trypsin digestion to form peptide sequences, and then they are delivered into the previously trained classifier. The classifier generates peptide detection probabilities. By summing all probabilities of peptides that were assigned to a protein leads to the Oi value of this particular protein. Reproduced, with permission, from ref[1].

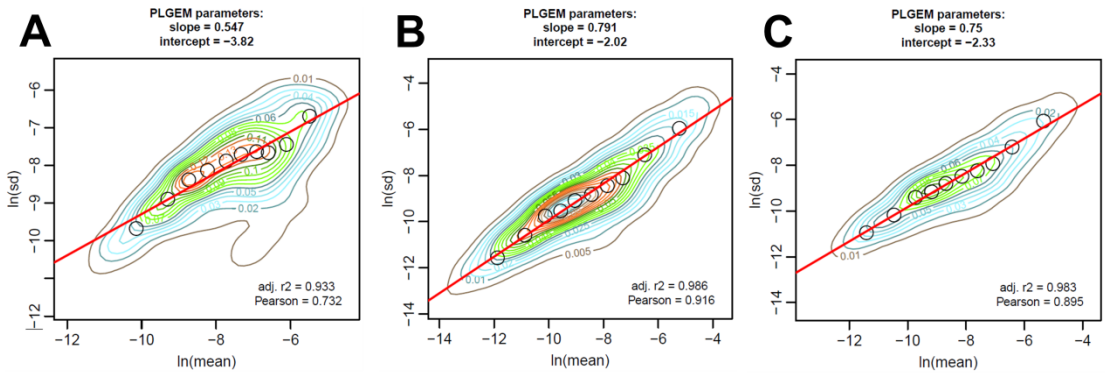

Figure S2. PLGEM diagnostic plot with variance-versus-mean trend line. The slope, r^2 and Pearson correlation coefficient demonstrate the goodness of fit to this model, all of which are closer to 1 the better. (A) APEX, (B) APEX-SI, and (C) APEX-SMT.

References

1. J. Braisted, S. Kuntumalla, C. Vogel, E. Marcotte, A. Rodrigues, R. Wang, S.T. Huang, E. Ferlanti, A. Saeed, R. Fleischmann, S. Peterson, R. Pieper, *BMC Bioinformatics*, 2008, **9**, 529.