CALCULATING ABSOLUTE PROTEIN ABUNDANCE FROM MASS SPECTROMETRY BASED PROTEIN EXPRESSION DATA - SUPPLEMENTARY NOTES

Christine Vogel[1], Edward M. Marcotte[1]*

[1] Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, University of Texas at Austin, 2500 Speedway, MBB 3.210, Austin, TX 78712

*Corresponding author

Emails: cvogel@mail.utexas.edu; edward.marcotte@gmail.com

phone: +1 512 232 3919 fax: +1 512 471 2149

ABBREVIATIONS:

MS – mass spectrometry; No. - number

PEPTIDES SEQUENCE ATTRIBUTES AND THEIR SOURCE

The number and types of peptide sequence attributes is important for performance of the training/testing of peptide MS detectability. Except for length, all amino acid attributes and their descriptions originate from AAindex (http://www.genome.jp/aaindex/). For attributes 42 to 66, both total (sum) and average values along sequence are included in the description of peptide properties (.arf and .arff files). We tested use of 2, 22, 42, 58 and 66 attributes (**Table S2**).

Table S1. Attributes

No. (cumulative) of attributes	Type / Descriptions	Source (reference number in AAindex)	Comment
	Length		
2	Molecular weight (Fasman, 1976)	FASG760101	Strongly correlated with Length
22	Relative amino acid frequencies	Instances of type of amino acid in sequence divided by length	
42	Absolute amino acid frequencies	Instances of type of amino acid in sequence	Correlated with Length
	Normalized frequency of alpha-helix (Chou-Fasman, 1978b)	CHOP780201	Secondary structure
	Normalized frequency of beta-sheet (Chou-Fasman, 1978b)	CHOP780202	Secondary structure
	Normalized frequency of beta-turn (Chou-Fasman, 1978b)	CHOP780203	Secondary structure
	Propensity to be buried inside (Wertz-Scheraga, 1978)	WERD780101	Main attribute for MUDPIT-ESI identified by Mallick et al. [1]
	Isoelectric point (Zimmerman et al., 1968)	ZIMJ680104	Main attribute for MUDPIT-ESI identified by Mallick et al. [1]
	Net charge (Klein et al., 1984)	KLEP840101	Main attribute for MUDPIT-ESI identified by Mallick et al. [1]
	Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)	EISD860102	Main attribute for MUDPIT-ESI identified by Mallick et al. [1]
58	Positive charge (Fauchere et al., 1988)	FAUJ880111	Main attribute for MUDPIT-ESI identified by Mallick et al. [1]
	Normalized flexibility parameters (B-values), average (Vihinen et al., 1994)	VINM940101	Additional attribute
	Normalized van der Waals volume (Fauchere et al., 1988)	FAUJ880103	Additional attribute
	Apparent partition energies calculated from Chothia index (Guy, 1985); Amino acid side-chain partition energies and distribution of residues in soluble proteins	GUYH850105	Additional attribute
66	Transfer energy, organic solvent/water (Nozaki-Tanford, 1971)	NOZY710101	Additional attribute

PARAMETER SENSITIVITY OF CLASSIFIER TRAINING

When building a model of peptide MS detectability, several parameters influence model quality as well as computation time. Model quality (performance) can be assessed by recall/precision (or ROC) plots (see **Figure S3**) or the F-measure (see main text). We modified the following parameters when optimizing model performance:

- Numbers and types of attributes. The more attributes are included, the better model performance, but also the larger is computation time. Performance (F-measure) improves significantly when moving from 2 attributes to 22 attributes or even more. For the final model, we used all 66 attributes.
    - 66 attributes: all attributes in **Table S1**
    - 58 attributes: similar to 66 attributes without Normalized flexibility parameters, Normalized van der Waals volume, Apparent partition energies, Transfer energy (solvent/water)
    - 42 attributes: length, molecular weight, relative and absolute amino acid frequencies
    - 22 attributes: length, molecular weight, relative amino acid frequencies (similar to original APEX paper[2])
    - 2 attributes: length, molecular weight
- Cost-sensitive training (or not). Model building largely fails without CostSensitive training or with inversed cost matrix (F-measure very low). Inversing the cost-matrix is a useful test for proper setup of the calculations.
- Selection of training set (proteins of high identification probability and/or with high total spectral count; peptides of high identification probability and/or with high spectral counts). For our final models, we selected a training set based on high protein identification probability (1.00) and high total spectral counts (>200 for LTQ-OrbiTrap and >50 for LCQ). These cutoffs are very likely to be dataset- and machine-dependent. However, as can be seen in the table, performance (F-measure) is insensitive to the number of proteins selected as long as the number is within the same range of 200 or 50 proteins, respectively.
- Memory allocation during WEKA use which influences computation time. We typically use 1800 MB, and building the final model for the LTQ-OrbiTrap (row in bold print) took ~4min.
- Type of mass spectrometer (ThermoFinnigan Surveyor/DecaXP+ iontrap (LCQ), ThermoFinnigan LTQ-OrbiTrap (ORBI)). Computation times and files are larger for the LTQ-OrbiTrap as well as spectral counts.
- Inclusion of degenerate proteins (protein groups of ambiguous identification) and degenerate peptides (peptides mapping to several proteins). We exclude both degenerate proteins and peptides.

In the table, rows in **bold** describe parameter settings of the final, best-performing models which are saved and used for prediction.

Table S2. Performance of classifier training

Min. protein identif. probability	Min. protein total spectral count per protein	Min. peptide identif. probability	Min. peptide spectral count per protein	No. of proteins in training set	No. of *observed peptides* in training set	Fraction	No. of *non-observed peptides* in training set	No. of attributes	F-measure of performance of class 1 (*Observed peptides*)	Comments	Secs taken to build model (without time for cross-validation)
ORBITRAP											
1	0	1	10	412	1447	0.02	59798	58	0.358		1298.24
1	0	1	30	102	258	0.02	12048	58	0.03	No cost matrix	113.11
1	0	1	30	102	258	0.02	12048	58	0.251		155.4
1	100	0	0	181	2373	0.08	26701	58	0.581		514.63
1	**200**	**0**	**0**	**89**	**1331**	**0.09**	**13279**	**66**	**0.614**		**238.31**
1	200	0	0	89	1331	0.09	13279	58	0.604		208.99
1	200	0	0	89	1331	0.09	13279	58	0	Inverse cost matrix	139.04
1	200	0	0	89	1331	0.09	13279	58	0.529	No cost matrix	200.87
1	200	0	0	89	1331	0.09	13279	42	0.591		168.59
1	200	0	0	89	1331	0.09	13279	22	0.578		103.72
1	200	0	0	89	1331	0.09	13279	2	0.215		35.84
1	300	0	0	53	809	0.10	7392	58	0.594		106.91
1	400	0	0	30	486	0.09	4755	58	0.533		58.23
LCQ											
1	25	0	0	95	710	0.05	13359	66	0.461		217.95
1	**50**	**0**	**0**	**50**	**460**	**0.06**	**6770**	**66**	**0.532**		**98.4**
1	75	0	0	34	361	0.07	4767	66	0.517		80.76

COMPARISON WITH PEPTIDE MS DETECTABILITY PREDICTIONS BY MALLICK ET AL.

For MUDPIT-ESI experiments, we mapped the peptide MS-detectability probabilities calculated by Mallick et al. [1] to those from our predictions (on LTQ-OrbiTrap, 66 attributes). Our probabilities compare very favorably to those by Mallick et al. – they are clearly shifted to higher values compared to a random sample of probabilities.

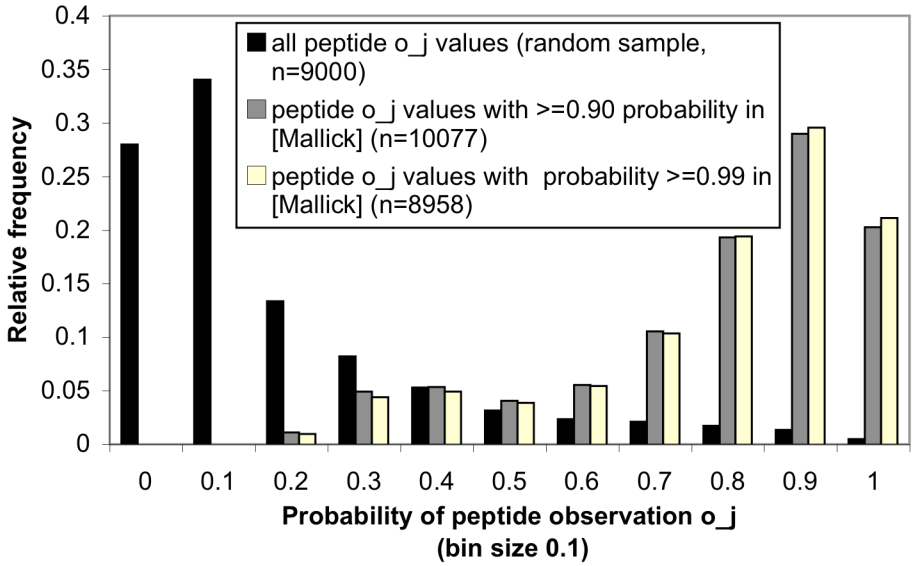


Figure S1. Comparison of predicted peptide MS detectability to Mallick et al.'s proteotytic peptide probabilities [1]

COMPARISON OF OI-VALUES FROM DIFFERENT MASS SPECTROMETERS

An LCQ mass spectrometer is much less sensitive than an LTQ-OrbiTrap, thus it is unsurprising that O_i values (expected unique number of peptides to be observed) for a given protein is lower on the LCQ than on the LTQ-OrbiTrap – however, since in both instruments the same ionization technique is used, O_i values correlate well.

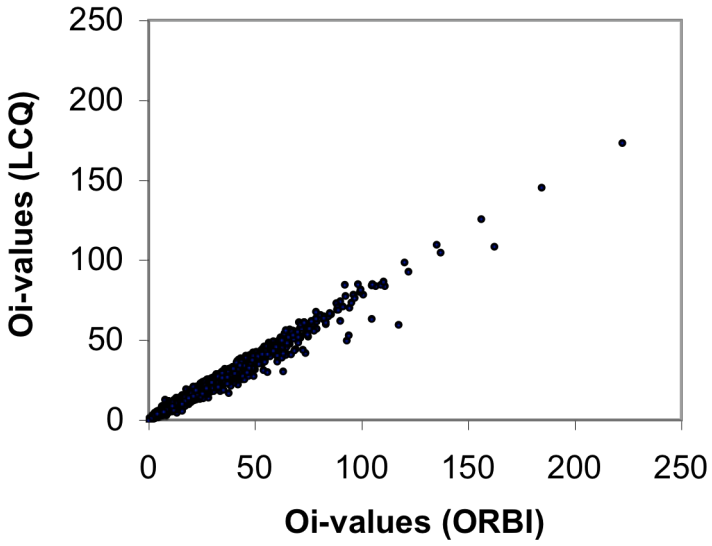
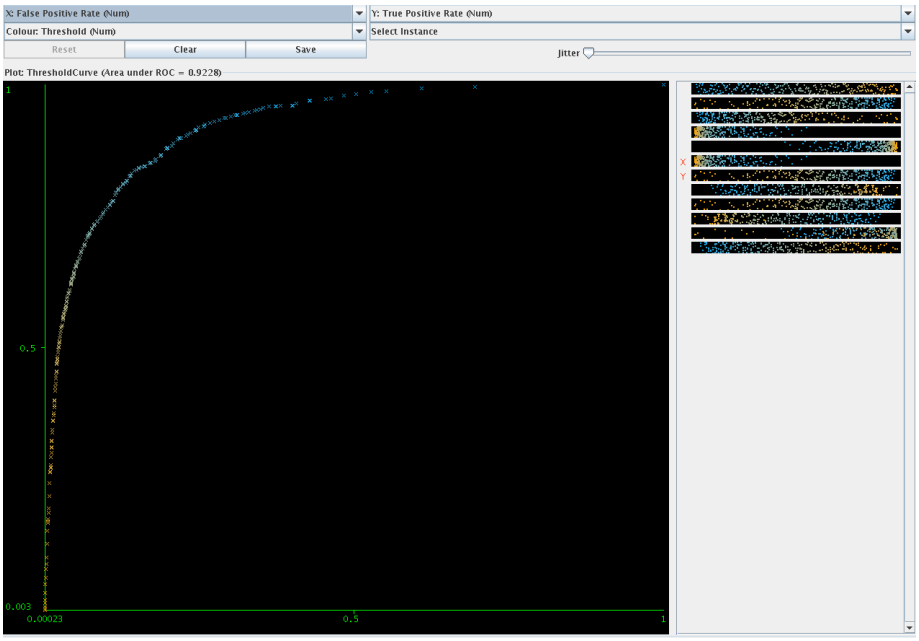


Figure S2. Oi values from ThermoFinnegan Surveyor/DecaPlus (LCQ) versus those from ThermoFinnegan LTQ-OrbiTrap (ORBI)

Data collected from yeast (cell lysate) grown in rich medium (YPD); the samples were independently prepared for analysis on the two machines. Tryptic peptides of 89 and 50 well identified proteins analyzed on LTQ-OrbiTrap (ORBI) and LCQ, respectively, were extracted based on their high protein identification probability (1.00) and high total spectral count (>200 and >50, respectively) and used for training. In both models, 66 attributes (**Table S1**) were used to describe peptide properties.

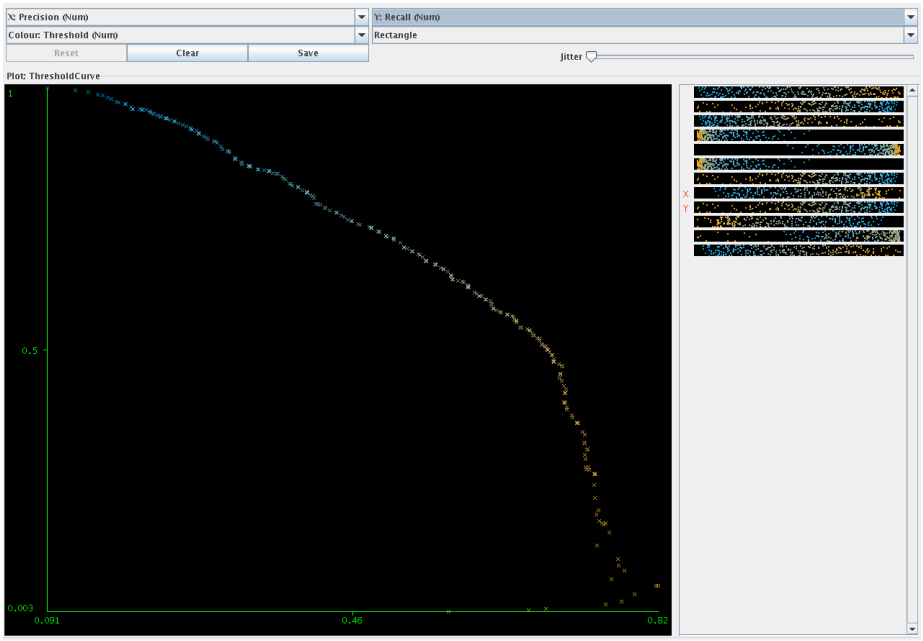
ROC PLOT AND PRECISION-RECALL-CURVE FOR LTQ-ORBITRAP MODEL

Tryptic peptides of 89 proteins analyzed on an LTQ-OrbiTrap were extracted based on their high protein identification probability (1.00) and high total spectral count (>200 and >50, respectively) and used for model training. Each peptide was described by 66 sequence attributes.



True positive rate

False positive rate



Recall

Precision

Figure S3. ROC and recall-precision plot of prediction of MS detectability for 89 proteins analyzed on LTQ-OrbiTrap

EXAMPLE OF Z-SCORE CALCULATIONS: YEAST GROWN IN MINIMAL VS. RICH MEDIUM

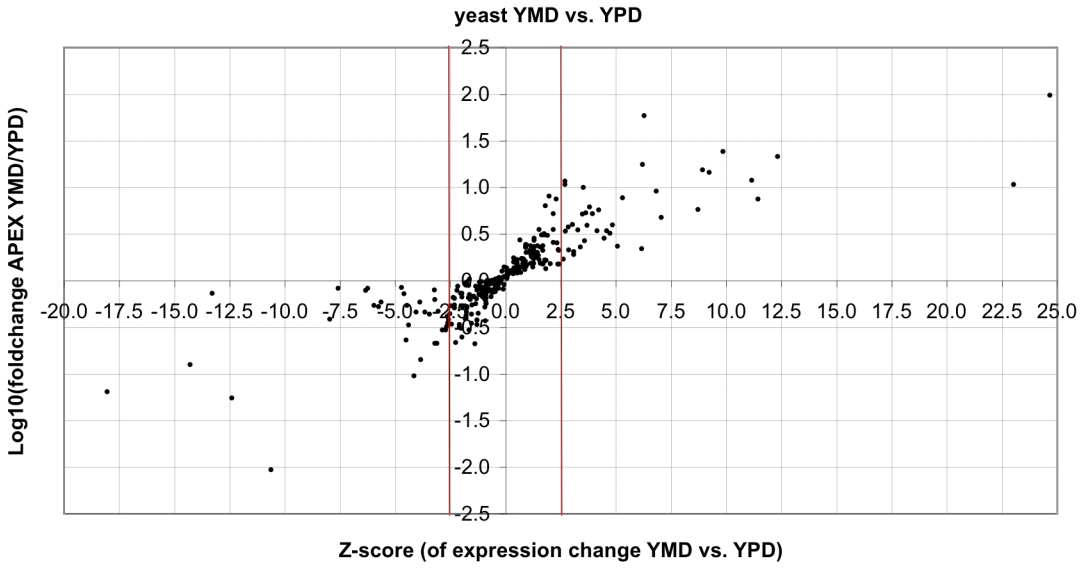


Figure S4. Z-score versus log10(fold-change APEX YMD/YPD)

Function analysis was performed using FuncAssociate (http://llama.med.harvard.edu/cgi/func/funcassociate) with the set of all genes detected in the MS/MS experiment as background. The data for the Z-score analysis is available at the Supplementary website, i.e. in the file yeast_YMD_YPD.zscore.gz at http://www.marcottelab.org/APEX_Protocol/Zscore_Yeast_YMD_vs_YPD/. We ignored all degenerate proteins (with ambiguous identification) and, for this plot, all proteins which are only measure in one of the samples.

Table S3. Function analysis of significantly up- or down-regulated genes

We analyzed sets of genes with Z<-2.58 (significant down-regulation in YMD vs. YPD, P-value<0.01) and Z>2.58 (significant up-regulation in YMD vs. YPD, P-value<0.01) for over- and under-represented functions. The genes are those within the red boundaries marked in **Figure S4**. There are no significantly over- or underrepresented functions for the first set of genes (P-value<0.05). The output has been copy-and-pasted from the FuncAssociate website. As expected, genes up-regulated in minimal medium (YMD) are enriched for functions in small molecule biosynthesis and amino acid biosynthesis.

Z<-2.58

OVERREPRESENTED ATTRIBUTES

Rank	N	X	LOD	P	P-adj	GO Attribute
none						

UNDERREPRESENTED ATTRIBUTES

Rank	N	X	LOD	P	P-adj	GO Attribute
none						

Z>2.58

OVERREPRESENTED ATTRIBUTES

Rank	N	X	LOD	P	P-adj	GO Attribute
1	31	55	1.101	2.30E-16	<0.001	0008652: amino acid biosynthesis
2	31	57	1.066	9.40E-16	<0.001	0044271: nitrogen compound biosynthesis
3	31	57	1.066	9.40E-16	<0.001	0009309: amine biosynthesis
4	46	124	0.836	2.10E-15	<0.001	0019752: carboxylic acid metabolism
5	46	124	0.836	2.10E-15	<0.001	0006082: organic acid metabolism
6	39	93	0.886	6.00E-15	<0.001	0006520: amino acid metabolism
7	41	103	0.857	8.40E-15	<0.001	0006807: nitrogen compound metabolism
8	40	100	0.855	1.70E-14	<0.001	0009308: amine metabolism
9	39	96	0.861	2.30E-14	<0.001	0006519: amino acid and derivative metabolism
10	12	15	1.452	2.50E-09	<0.001	0009084: glutamine family amino acid biosynthesis
11	73	393	0.513	1.40E-07	<0.001	0003824: catalytic activity/enzyme activity
12	11	16	1.213	1.70E-07	<0.001	0006790: sulfur metabolism/sulphur metabolism
13	12	20	1.063	4.00E-07	0.001	0009064: glutamine family amino acid metabolism
14	46	208	0.456	2.40E-06	0.002	0044249: cellular biosynthesis
15	12	23	0.93	3.10E-06	0.002	0009066: aspartate family amino acid metabolism

16	6	6	1.986	3.30E-06	0.002	0006526: arginine biosynthesis
17	23	73	0.595	4.70E-06	0.002	0016491: oxidoreductase activity/redox activity
18	9	14	1.121	6.00E-06	0.002	0000096: sulfur amino acid metabolism/sulphur amino acid metabolism
19	47	221	0.433	6.20E-06	0.002	0009058: biosynthesis/anabolism
20	8	12	1.155	1.40E-05	0.002	0006555: methionine metabolism
21	6	7	1.508	2.00E-05	0.003	0000051: urea cycle intermediate metabolism
22	6	7	1.508	2.00E-05	0.003	0006525: arginine metabolism
23	9	16	0.985	2.80E-05	0.005	0009069: serine family amino acid metabolism
24	83	541	0.42	0.00013	0.016	0008152: metabolism/metabolic process
25	11	27	0.729	0.00015	0.022	0016614: oxidoreductase activity, acting on CH-OH group of donors
26	81	527	0.393	0.00022	0.03	0044237: cellular metabolism

UNDERREPRESENTED ATTRIBUTES

Rank	N	X	LOD	P	P-adj	GO Attribute
1	10	201	-0.509	5.60E-05	0.009	0019538: protein metabolism/protein metabolism and modification
2	13	225	-0.449	0.00012	0.015	0043283: biopolymer metabolism
3	10	192	-0.481	0.00015	0.022	0044267: cellular protein metabolism

SUPPLEMENTARY WEBSITE

The Supplementary website is at http://www.marcottelab.org/APEX_Protocol/ and contains:

- Perl scripts for all steps of file parsing
- input files (ProteinProphet output files)
- training data and training results
- intermediate data files
- prediction data and results
- O_i-values for *E. coli*, yeast and human
- example Z-score calculation (yeast grown in YMD and YPD)

REFERENCES

1. Mallick, P. et al. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol* **25**, 125-131 (2007).

2. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E.M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* **25**, 117-124 (2007).