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

Christine Vogel¹, Edward M. Marcotte¹*

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

<table>
<thead>
<tr>
<th>No. (cumulative) of attributes</th>
<th>Type / Descriptions</th>
<th>Source (reference number in AAindex)</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Length</td>
<td>FASG760101</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Molecular weight (Fasman, 1976)</td>
<td>FASG760101</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>Relative amino acid frequencies</td>
<td>Instances of type of amino acid in sequence divided by length</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>Absolute amino acid frequencies</td>
<td>Instances of type of amino acid in sequence</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Normalized frequency of alpha-helix (Chou-Fasman, 1978b)</td>
<td>CHOP780201</td>
<td>Secondary structure</td>
</tr>
<tr>
<td></td>
<td>Normalized frequency of beta-sheet (Chou-Fasman, 1978b)</td>
<td>CHOP780202</td>
<td>Secondary structure</td>
</tr>
<tr>
<td></td>
<td>Normalized frequency of beta-turn (Chou-Fasman, 1978b)</td>
<td>CHOP780203</td>
<td>Secondary structure</td>
</tr>
<tr>
<td></td>
<td>Propensity to be buried inside (Wertz-Scheraga, 1978)</td>
<td>WERD780101</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Isoelectric point (Zimmerman et al., 1968)</td>
<td>ZIMJ680104</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Net charge (Klein et al., 1984)</td>
<td>KLEP840101</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)</td>
<td>EISD860102</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>Positive charge (Fauchere et al., 1988)</td>
<td>FAUJ880111</td>
<td>Main attribute for MUDPIT-ESI identified by Mallick et al.</td>
</tr>
<tr>
<td></td>
<td>Normalized flexibility parameters (B-values), average (Vihinen et al., 1994)</td>
<td>VINM940101</td>
<td>Additional attribute</td>
</tr>
<tr>
<td></td>
<td>Normalized van der Waals volume (Fauchere et al., 1988)</td>
<td>FAUJ880103</td>
<td>Additional attribute</td>
</tr>
<tr>
<td></td>
<td>Apparent partition energies calculated from Chothia index (Guy, 1985); Amino acid side-chain partition energies and distribution of residues in soluble proteins</td>
<td>GUYH850105</td>
<td>Additional attribute</td>
</tr>
<tr>
<td>66</td>
<td>Transfer energy, organic solvent/water (Nozaki-Tanford, 1971)</td>
<td>NOZY710101</td>
<td>Additional attribute</td>
</tr>
</tbody>
</table>
**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 paper²)
  - 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

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

![Bar chart showing comparison of predicted peptide MS detectability to Mallick et al.'s proteolytic peptide probabilities](image)

**Figure S1.** Comparison of predicted peptide MS detectability to Mallick et al.’s proteolytic 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.

![Image](image_url)

**Figure S2.** Oi values from ThermoFinnigan Surveyor/DecaPlus (LCQ) versus those from ThermoFinnigan 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.

![ROC plot](image1)

**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**

<table>
<thead>
<tr>
<th>Rank</th>
<th>N</th>
<th>X</th>
<th>LOD</th>
<th>P</th>
<th>P-adj</th>
<th>GO Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>none</td>
</tr>
</tbody>
</table>

**UNDERREPRESENTED ATTRIBUTES**

<table>
<thead>
<tr>
<th>Rank</th>
<th>N</th>
<th>X</th>
<th>LOD</th>
<th>P</th>
<th>P-adj</th>
<th>GO Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>none</td>
</tr>
</tbody>
</table>

### Z>2.58

**OVERREPRESENTED ATTRIBUTES**

<table>
<thead>
<tr>
<th>Rank</th>
<th>N</th>
<th>X</th>
<th>LOD</th>
<th>P</th>
<th>P-adj</th>
<th>GO Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31</td>
<td>55</td>
<td>1.101</td>
<td>2.30E-16</td>
<td>&lt;0.001</td>
<td>0008652: amino acid biosynthesis</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>57</td>
<td>1.066</td>
<td>9.40E-16</td>
<td>&lt;0.001</td>
<td>0044271: nitrogen compound biosynthesis</td>
</tr>
<tr>
<td>3</td>
<td>31</td>
<td>57</td>
<td>1.066</td>
<td>9.40E-16</td>
<td>&lt;0.001</td>
<td>0009309: amine biosynthesis</td>
</tr>
<tr>
<td>4</td>
<td>46</td>
<td>124</td>
<td>0.836</td>
<td>2.10E-15</td>
<td>&lt;0.001</td>
<td>0019752: carboxylic acid metabolism</td>
</tr>
<tr>
<td>5</td>
<td>46</td>
<td>124</td>
<td>0.836</td>
<td>2.10E-15</td>
<td>&lt;0.001</td>
<td>0006082: organic acid metabolism</td>
</tr>
<tr>
<td>6</td>
<td>39</td>
<td>93</td>
<td>0.886</td>
<td>6.00E-15</td>
<td>&lt;0.001</td>
<td>0006520: amino acid metabolism</td>
</tr>
<tr>
<td>7</td>
<td>41</td>
<td>103</td>
<td>0.857</td>
<td>8.40E-15</td>
<td>&lt;0.001</td>
<td>0006807: nitrogen compound metabolism</td>
</tr>
<tr>
<td>8</td>
<td>40</td>
<td>100</td>
<td>0.855</td>
<td>1.70E-14</td>
<td>&lt;0.001</td>
<td>0009308: amine metabolism</td>
</tr>
<tr>
<td>9</td>
<td>39</td>
<td>96</td>
<td>0.861</td>
<td>2.30E-14</td>
<td>&lt;0.001</td>
<td>0006519: amino acid and derivative metabolism</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>15</td>
<td>1.452</td>
<td>2.50E-09</td>
<td>&lt;0.001</td>
<td>0009084: glutamine family amino acid biosynthesis</td>
</tr>
<tr>
<td>11</td>
<td>73</td>
<td>393</td>
<td>0.513</td>
<td>1.40E-07</td>
<td>&lt;0.001</td>
<td>0003824: catalytic activity/enzyme activity</td>
</tr>
<tr>
<td>12</td>
<td>11</td>
<td>16</td>
<td>1.213</td>
<td>1.70E-07</td>
<td>&lt;0.001</td>
<td>0006790: sulfur metabolism/sulphur metabolism</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>20</td>
<td>1.063</td>
<td>4.00E-07</td>
<td>0.001</td>
<td>0009064: glutamine family amino acid metabolism</td>
</tr>
<tr>
<td>14</td>
<td>46</td>
<td>208</td>
<td>0.456</td>
<td>2.40E-06</td>
<td>0.002</td>
<td>0044249: cellular biosynthesis</td>
</tr>
<tr>
<td>15</td>
<td>12</td>
<td>23</td>
<td>0.93</td>
<td>3.10E-06</td>
<td>0.002</td>
<td>0009066: aspartate family amino acid metabolism</td>
</tr>
<tr>
<td>Rank</td>
<td>N</td>
<td>X</td>
<td>LOD</td>
<td>P</td>
<td>P-adj</td>
<td>GO Attribute</td>
</tr>
<tr>
<td>------</td>
<td>----</td>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>201</td>
<td>-0.509</td>
<td>5.60E-05</td>
<td>0.009</td>
<td>0019538: protein metabolism/protein metabolism and modification</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>225</td>
<td>-0.449</td>
<td>0.00012</td>
<td>0.015</td>
<td>0043283: biopolymer metabolism</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>192</td>
<td>-0.481</td>
<td>0.00015</td>
<td>0.022</td>
<td>0044267: cellular protein metabolism</td>
</tr>
</tbody>
</table>

UNDERREPRESENTED ATTRIBUTES

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

UNDER REPRESENTED ATTRIBUTES

<table>
<thead>
<tr>
<th>Rank</th>
<th>N</th>
<th>X</th>
<th>LOD</th>
<th>P</th>
<th>P-adj</th>
<th>GO Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>201</td>
<td>-0.509</td>
<td>5.60E-05</td>
<td>0.009</td>
<td>0019538: protein metabolism/protein metabolism and modification</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>225</td>
<td>-0.449</td>
<td>0.00012</td>
<td>0.015</td>
<td>0043283: biopolymer metabolism</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>192</td>
<td>-0.481</td>
<td>0.00015</td>
<td>0.022</td>
<td>0044267: cellular protein metabolism</td>
</tr>
</tbody>
</table>
SUPPLEMENTARY WEBSITE


- 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
- $O_\text{-}$-values for *E. coli*, yeast and human
- example Z-score calculation (yeast grown in YMD and YPD)

REFERENCES
