Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein*§

Yasushi Ishihama‡§¶, Yoshiya Oda‡, Tsuyoshi Tabata‡, Toshitaka Sato‡, Takeshi Nagasu‡, Juri Rappsilber§∥, and Matthias Mann§**

To estimate absolute protein contents in complex mixtures, we previously defined a protein abundance index (PAI) as the number of observed peptides divided by the number of observable peptides per protein (Rappsilber, J., Ryder, U., Lamond, A. I., and Mann, M. (2002) Large-scale proteomic analysis of the human spliceosome. Genome Res. 12, 1231–1245). Here we report that PAI values obtained at different concentrations of serum albumin show a linear relationship with the logarithm of protein concentration in LC-MS/MS experiments. This was also the case for 46 proteins in a mouse whole cell lysate. For absolute quantitation, PAI was converted to exponentially modified PAI (emPAI), equal to 10PAI minus one, which is proportional to protein content in a protein mixture. For the 46 proteins in the whole lysate, the deviation percentages of the emPAI-based abundances from the actual values were within 63% on average, similar or better than determination of abundance by protein staining. emPAI was applied to comprehensive protein expression analysis and to a comparison study between gene and protein expression in a human cancer cell line, HCT116. The values of emPAI are easily calculated and add important quantitative information to proteomic experiments; therefore we suggest that they should be reported in large scale proteomic identification projects. Molecular & Cellular Proteomics 4:1265–1272, 2005.

Proteomic LC-MS approaches combined with genome-annotated databases currently allow identification of thousands of proteins from complex mixtures (1). Approaches have also been developed for relative quantitation using stable isotope labeling (2–4). Recently not only comprehensive quantitation studies between two states (5, 6) but also protein-protein (7, 8), protein-peptide (9), and protein-drug (10) interaction analyses have been reported. So far, however, a comprehensive approach for determining protein concentrations in one sample has not been established. Protein concentrations are one of the most basic and important parameters in quantitative proteomics because the kinetics/dynamics of the cellular proteome is described in terms of changes in the concentrations of proteins in particular compartments. Biological experiments often require at least some information on protein abundance for correct interpretation. In the past, crude quantitative information could be drawn from the intensity of gel staining in comparison to a known amount of marker protein. However, in complex mixture analysis, individual proteins cannot be stained individually, and usually all information about protein abundance is lost. So far, isotope-labeled synthetic peptides have been used as internal standards for absolute quantitation of particular proteins of interest (11, 12). This approach is in principle applicable to comprehensive analysis but is hampered by the high cost of isotope-labeled peptides as well as the difficulty of quantitative digestion of proteins in-gel (13).

Even a single nano-LC-MS/MS analysis can easily generate a long list of identified proteins with the help of database searching, and additional information can be extracted, such as the hit rank in identification, the probability score, the number of identified peptides per protein, ion counts of identified peptides, LC retention times, and so on. Qualitatively some parameters, such as the hit rank, the score, and the number of peptides per protein (14), can be considered as indicators for protein abundance in the analyzed sample. Among them, the integrated ion counts of the peptides identifying each protein would be the most direct parameter to describe the abundance and has been used to compare protein expression in different states (15). However, a mass spectrometer is not as versatile as an absorbance detector because of limited linearity and possibly because of background and ionization suppression effects (16). Therefore, it is necessary to normalize these parameters to obtain at least approximate quantitative information. The first approach to achieve this, to our knowledge, was to use the number of peptides per protein normalized by the theoretical number of...
emPAI for Protein Abundance Estimation

peptides (so-called protein abundance index (PAI))\(^1\), and this was applied to human spliceosome complex analysis (17). PAI is superior to the number of identified peptides because it takes account of the fact that, for the same number of molecules, larger proteins and proteins with many peptides in the preferred mass range for mass spectrometry will generate more observed peptides. Independently Sanders et al. (18) developed a similar index. The number of peptides, spectra counts, or the total of the peptide probability scores in LC/LC-MS/MS analysis can also be used for relative quantitation (19–21). Here we further develop the PAI strategy to determine protein abundance from nano-LC-MS/MS experiments and present a modified form, emPAI, the exponential form of PAI minus one. In experiments with labeled complex mixtures, into which we spiked in synthetic peptides, we show emPAI to be roughly proportional to protein abundance.

MATERIALS AND METHODS

Preparation of Cell Lysate—RPMI 1640 medium (Invitrogen) containing \(^{13}\)C\(_6\)Leu (Cambridge Isotope Laboratories, Andover, MA) was prepared according to the SILAC protocol of Ong et al. (4). Mouse neuroblastoma neuro2a cells were cultured in this medium for \(^{13}\)C\(_6\)Leu labeling. Whole cells were lysed using ultrasonication in the presence of a protease inhibitor mixture (Roche Diagnostics). HCT116-C9 cells were grown in a normal RPMI 1640 culture medium as described previously (10). Whole proteins were extracted with 5 ml of M-PER (Pierce) containing protease inhibitor mixture and 5 mm dithiothreitol.

Preparation of Peptide Mixtures for LC-MS/MS—Proteins from cell lysates were dried and resuspended in 50 mm Tris-HCl buffer (pH 9.0) containing 8 mm urea. These mixtures were subsequently reduced, alkylated, and digested with Lys-C (Wako, Osaka, Japan) and trypsin (Promega, Madison, WI) as described previously (6). Digested solutions were acidified with TFA and desalted and concentrated using C\(_{18}\) StageTips (22), which were prepared by a fully automated instrument (Nikkoyo Technos, Tokyo, Japan) with Empore C\(_{18}\) disks (3M, St. Paul, MN). Peptide fractionation by strong cation exchange chromatography (SCX) was performed using SCX-StageTip with 0–500 mm five-step ammonium acetate salt elution (23), and the resultant fractions were desalted using C\(_{8}\) StageTips prior to LC-MS/MS analysis. Candidates for peptide synthesis containing at least one leucine and one tyrosine were selected, considering the sequences of tryptic peptides from proteins expressed in neuro2a cells. Peptides containing methionine and tryptophan were removed to avoid oxidation problems during sample preparation. In addition, peptides with double basic residues were removed, considering the frequency of missed cleavage by trypsin. The selected 54 peptides were synthesized using a Shimadzu PSSM-8 (Kyoto, Japan) with Fmoc (N-(9-fluorenylmethoxycarbonyl) chemistry and were purified by preparative HPLC. Amino acid analysis, peptide mass measurement, and HPLC-UV were carried out for purity and structure elucidation. A solution containing equal amounts of each peptide was spiked into the peptide mixtures from neuro2a cells. Three different amounts were spiked so that peak intensity ratios of unlabeled peptides to labeled peptides were between 0.2 and 5.

\[^{1}\]The abbreviations used are: PAI, protein abundance index; emPAI, exponentially modified protein abundance index; SILAC, stable isotope labeling with amino acids in cell culture; SCX, strong cation exchange chromatography; HSA, human serum albumin.

Nano-LC-MS/MS Analysis—All samples were analyzed by nano-LC-MS using a QSTAR Pulsar i (AB/MDSCIEX, Toronto, Canada), a Finnigan LCO Advantage (Thermoelectron, San Jose, CA) or a Finnigan LTQ (Thermoelectron) system equipped with a Shimadzu LC10A gradient pump and an HTC-PAL autosampler (CTC Analytics AG, Zwingen, Switzerland) equipped with Valco C2 valves with 150-\(\mu\)m ports. ReproSil C\(_{18}\) materials (3 \(\mu\)m, Dr. Maisch, Ammerbuch, Germany) were packed into a self-pulled needle (Nikkoyo) to prepare an analytical column needle with “stone-arch” frit (24). A Teflon-coated column holder (Nikkoyo) was mounted on an x-y-z nanospray interface (Proxeon, Odense, Denmark), and a Valco metal connector with a magnet was used to hold the column needle and to set the appropriate spray position. The injection volume was 2.5 \(\mu\)l and the flow rate was 250 nl/min after a tee splitter. The mobile phases consisted of A (0.5% acetic acid) and B (0.5% acetic acid and 80% acetonitrile). The three-step linear gradient of 5–10% B in 5 min, 10–30% in 60 min, 30–100% in 5 min, and 100% in 10 min was used throughout this study. A spray voltage of 2400 V was applied via the metal connector as described previously (24). For QSTAR experiments with the faster scan mode, MS scans were performed for 1 s to select three intense peaks, and subsequently three MS/MS scans were performed for 0.55 s each. An information-dependent acquisition function was active for 3 min to exclude the previously scanned parent ions. For the slower scan mode, four MS/MS scans (1.5 s each) per one MS scan (1 s) were performed. For LCQ and LTQ experiments, two MS/MS scans per one MS scan were performed in the automated gain control mode. The scan cycle was 1.19 s for one MS and 1.17 s for one MS/MS on average in LCQ and 0.17 s for one MS and 0.38 s for one MS/MS on average in LTQ. The scan range was m/z 350–1400 for QSTAR, LCQ, and LTQ.

Data Analysis—A Mascot version 1.9 database search engine (Matrix Sciences, London, UK) was used for protein identification against the Swiss-Prot protein database. The allowed number of missed cleavages was set to 1, and peptide scores to indicate identity were used for peptide identification without manual inspection of MS/MS spectra. MSQuant version 1.4a was customized for \(^{13}\)C\(_6\)Leu SILAC to determine the ion counts in chromatograms for absolute concentrations of proteins using known amounts of synthetic peptides. MSQuant is open source software developed by us and available at sourceforge.net.

Protein Abundance Determination—To calculate the number of observable peptides per protein, proteins were digested in silico, and the obtained peptides masses were compared with the scan range of the mass spectrometer. In addition, the expected retention times under our nano-LC conditions were calculated according to the procedure of Meek (25) and Sakamoto et al. (26) with our own coefficients based on ~1500 peptides. Peptides that were too hydrophilic or hydrophobic were eliminated. An in-house program was written in PHP to calculate the peptide number and was used to export all data to Microsoft Excel. The program is freely accessible at xome.hydra.mki.co.jp/8080/btt/common/Menu. Regarding the number of observed peptides per protein, three methods of counting were used, i.e. 1) counting unique parent ions, 2) counting unique sequences, and 3) counting unique sequences without partial modification and the overlap caused by missed tryptic cleavage. These numbers were exported from Mascot html files to Excel spreadsheets using the “Export All Peptides” function of MSQuant software.

The PAI is defined as

\[
P_{\text{PAI}} = \frac{N_{\text{obsd}}}{N_{\text{obsbl}}} \tag{1}
\]

where \(N_{\text{obsd}}\) and \(N_{\text{obsbl}}\) are the number of observed peptides per
emPAI for Protein Abundance Estimation

protein and the number of observable peptides per protein, respectively (17). The emPAI is defined as follows.

\[
\text{emPAI} = 10^{\text{PAI}} - 1
\]  
(Eq. 2)

Thus, the protein contents in molar and weight fraction percentages are described as

\[
\text{Protein content (mol \%) = \frac{\text{emPAI}}{\sum (\text{emPAI})} \times 100}
\]  
(Eq. 3)

\[
\text{Protein content (weight \%) = \frac{\text{emPAI} \times M_i}{\sum (\text{emPAI} \times M_i)} \times 100}
\]  
(Eq. 4)

where \(M_i\) is the molecular weight of the protein, and \(\sum (\text{emPAI})\) is the summation of emPAI values for all identified proteins. The entire procedure for emPAI calculation is shown in Supplemental Sheet 1.

To evaluate the accuracy of the parameters, a deviation factor was defined as

\[
\text{Deviation factor} = \frac{\text{Value}_{\text{measured}}}{\text{Value}_{\text{estimated}}}
\]  
(Eq. 5)

where measured values are larger than estimated values or

\[
\text{Deviation factor} = \frac{\text{Value}_{\text{estimated}}}{\text{Value}_{\text{measured}}}
\]  
(Eq. 6)

where estimated values are larger than measured values.

**RESULTS AND DISCUSSION**

**The Number of Identified Peptides from a Single Protein at Different Concentrations**—Different amounts of human serum albumin (HSA) tryptic peptides were analyzed by nano-LC-ESI-MS/MS, and the number of identified peptides was counted. As shown in Fig. 1A, both peak area and the number of identified peptides with unique parent ions increased as the injection amount increased, although both curves saturated at
emPAI for Protein Abundance Estimation

larger amounts of HSA close to 1 pmol. However, even in the region where the peak area is linear, the number of peptides does not have a linear relationship to the protein amount. Interestingly the number of peptides shows a linear relationship to the logarithm of the injected amount from 3 to 500 fmol (Fig. 1B). A similar result was obtained on an LCQ with the slower scan cycle (see “Materials and Methods”). This finding indicates that each peak was well separated in time and that the influence of “random sampling” caused by the slower scan could be neglected under this condition. In this case, three ways were used to count peptides: 1) all parent ions including different charge states from the same peptide sequences, 2) all peptides excluding different charge states and partial modifications such as methionine oxidation, and 3) peptides with unique sequences excluding peptides overlapped by missed tryptic cleavage sites. Fig. 1B shows that the number of peptides based on unique parent ions (Fig. 1B, squares, and 1) above) shows the best correlation with the logarithm of protein abundance. We believe that these results are not due to the particular conditions used but are a more general phenomenon. Recently two groups independently presented similar curves relating the number of peptides to the concentration of proteins (19, 27). Although neither of them analyzed the logarithmic relationship, it appears to us that their data are also consistent with a linear relationship between the logarithm of protein concentration and the number of peptides. At present, it is not clear why the logarithm of protein concentration correlates with the number of observed peptides, and in any case this relationship is likely to be due to a combination of processes and probably holds only approximately. In any case, it is a common experience that the mass spectrometric peptide signals from the digestion of a protein are vastly different. For example, to substantially increase sequence coverage of a protein often requires orders of magnitude large protein amounts, and conversely dilution by small factors often does not decrease sequence coverage very much.

PAI of 46 Proteins in Complex Mixture Solutions—To test performance of the PAI index in complex mixtures, we next investigated known amounts of 54 proteins in a whole cell lysate. Tryptic peptides from mouse neuroblastoma neuro2a cells SILAC-labeled with \(^{[13]C_6}\)Leu (4) were measured by a single LC-MS/MS run with the QSTAR instrument, and 336 peptides were identified based on 1462 peptides. For accurate absolute quantitation, we spiked 54 synthetic peptides containing \(^{[13]C_6}\)Leu into this sample solution, one for each protein, and quantified the corresponding tryptic peptides containing \(^{[13]C_6}\)Leu. Eight peptides were not quantified because they resulted in overlapping peaks in the extracted ion current chromatograms. Together 46 proteins ranging in molecular mass from 13 to 193 kDa were quantified in the range from 30 fmol to 1.8 pmol/µl in the sample solution as listed in Table I.

In complex protein mixtures, two additional factors should be considered. One is the influence of protein size on the number of peptides. Generally larger proteins generate more detectable peptides. Therefore, observable peptides were used for normalization as done previously except that we used the predicted peptide retention time as an additional filter. The other factor is the mixture complexity. A very large number of peptides exist in total cell lysate, and the number of observed peptides could to some extent be influenced by the random selection for MS/MS events, ion suppression effects, and saturation of the MS analyzer and/or the detector. Nevertheless Fig. 1C shows that there is still a linear relationship between log[protein] and the number of observed peptides normalized by the number of observable peptides per protein even when different proteins were plotted into one graph. Compared with other parameters, PAI correlated most highly with logarithm of protein amount (Fig. 1C, r = 0.89, deviation factor (average ± S.D.) = 1.6 ± 0.5) followed by number of peptides divided by protein M, (Fig. 1D, r = 0.84, deviation factor = 1.8 ± 0.8), a measure similar to PAI except that it ignores how well the peptide sequence can generate tryptic peptides in the correct mass range for mass spectrometry. Commonly used proxies for protein abundance such as Mascot score and the number of peptides correlate much worse with protein abundance (r = 0.72, deviation factor = 2.7 ± 2.4 and r = 0.71, deviation factor = 2.7 ± 2.6, respectively).

Absolute Quantitation Using emPAI—Although PAI can estimate the abundance relationship between proteins, it cannot express the molar fraction directly. Therefore, we derived a new parameter, emPAI, that is the exponential form of PAI minus 1 (Equation 2) and that is directly proportional to protein content as shown in Fig. 1E. To calculate the absolute concentrations, total protein amounts were measured as weight by BCA assay, and the weight fractions of 46 proteins among 336 neuro2a proteins were calculated using Equation 4. As shown in Fig. 1F, the emPAI-based concentrations were highly consistent with the actual values (y = 0.973x, r = 0.93), and the deviation factors ranged from 1.03 to 4.98 with an average of 1.74 ± 0.79. The outlier in (x, y) = (10.6, 2.13) is clathrin heavy chain (CLH_RAT). Mouse clathrin is not in the current Swiss-Prot but in TrEMBL (Q68FD5_MOUSE), which was not used for protein identification. Q68FD5_MOUSE is not identical in sequence to CLH_RAT. It is possible that the number of observed peptides would increase using Q68FD5_MOUSE or other sequences instead of CLH_RAT, although Q68FD5_MOUSE needs more validation for Swiss-Prot entry. Note that these measures of confidence compare favorably with protein abundance by comparative gel staining and indeed with the Bradford assay itself used here to measure total protein amount (28). Furthermore just as there are proteins known to stain well, the emPAI of certain proteins could also be adjusted in the future. In any case, the emPAI approach seems to provide a reasonably accurate estimate for comprehensive absolute quantitation.

Dependence of emPAI on Experimental Conditions—In this experiment, we used the fast MS and MS/MS cycle time on...
the QSTAR to maximize the number of MS/MS events. When a slower cycle was used, the deviation from the linear relationship between emPAI and the protein concentrations increased (Fig. 2, A and B) due to the random sampling effects mentioned above. This effect was more pronounced when an LCQ ion trap instrument was used (Fig. 2C) presumably because the limited trap capacity results in a biased peak selection for more abundant proteins, and indeed a larger deviation was observed for more abundant proteins. We used an LTQ, a linear ion trap instrument that has a higher capacity and a faster scan time when compared with the LCQ (29, 30), to evaluate the influence of the cycle time. The use of LTQ data improved the accuracy of emPAI in comparison to LCQ data (Fig. 2D). However, the faster scan cycle of LTQ com-

| Table I |
| 46 proteins identified and quantified in mouse neuro2a cells |

<table>
<thead>
<tr>
<th>Swiss-Prot accession no.</th>
<th>Protein name</th>
<th>Mascot hit no.</th>
<th>Mr</th>
<th>Protein concentration*</th>
<th>Number of observed peptides</th>
<th>Mascot score</th>
<th>PAI</th>
<th>emPAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>P19378</td>
<td>Heat shock cognate 71-kDa protein</td>
<td>1</td>
<td>70,989</td>
<td>822</td>
<td>29</td>
<td>1235</td>
<td>0.88</td>
<td>6.56</td>
</tr>
<tr>
<td>P07901</td>
<td>Heat shock protein HSP 90-α</td>
<td>2</td>
<td>85,003</td>
<td>940</td>
<td>31</td>
<td>1047</td>
<td>0.86</td>
<td>6.26</td>
</tr>
<tr>
<td>P20152</td>
<td>Vimentin</td>
<td>5</td>
<td>53,581</td>
<td>336</td>
<td>27</td>
<td>1060</td>
<td>0.82</td>
<td>5.58</td>
</tr>
<tr>
<td>P58252</td>
<td>Elongation factor 2 (EF-2)</td>
<td>6</td>
<td>96,091</td>
<td>118</td>
<td>24</td>
<td>830</td>
<td>0.53</td>
<td>2.41</td>
</tr>
<tr>
<td>Q03265</td>
<td>ATP synthase α chain, mitochondrial precursor</td>
<td>8</td>
<td>59,830</td>
<td>149</td>
<td>18</td>
<td>635</td>
<td>0.56</td>
<td>2.65</td>
</tr>
<tr>
<td>P17182</td>
<td>α enolase</td>
<td>9</td>
<td>47,322</td>
<td>596</td>
<td>21</td>
<td>828</td>
<td>0.88</td>
<td>6.50</td>
</tr>
<tr>
<td>P15331</td>
<td>Peripherin</td>
<td>10</td>
<td>54,349</td>
<td>84</td>
<td>13</td>
<td>556</td>
<td>0.39</td>
<td>1.48</td>
</tr>
<tr>
<td>P48975</td>
<td>Actin, cytoplasmic 1 (β-actin)</td>
<td>11</td>
<td>42,053</td>
<td>1206</td>
<td>22</td>
<td>894</td>
<td>1.22</td>
<td>15.68</td>
</tr>
<tr>
<td>P05213</td>
<td>Tubulin α-2 chain (α-tubulin 2)</td>
<td>14</td>
<td>50,818</td>
<td>1782</td>
<td>24</td>
<td>862</td>
<td>1.14</td>
<td>12.89</td>
</tr>
<tr>
<td>P52480</td>
<td>Pyruvate kinase, M2 isozyme</td>
<td>16</td>
<td>58,289</td>
<td>216</td>
<td>17</td>
<td>643</td>
<td>0.49</td>
<td>2.06</td>
</tr>
<tr>
<td>P20001</td>
<td>Elongation factor 1-α 1</td>
<td>24</td>
<td>50,424</td>
<td>870</td>
<td>17</td>
<td>647</td>
<td>1.00</td>
<td>9.00</td>
</tr>
<tr>
<td>P06113</td>
<td>Endoplasmic precursor</td>
<td>25</td>
<td>92,703</td>
<td>90</td>
<td>10</td>
<td>379</td>
<td>0.23</td>
<td>0.71</td>
</tr>
<tr>
<td>O35501</td>
<td>Stress-70 protein, mitochondrial precursor</td>
<td>27</td>
<td>73,970</td>
<td>195</td>
<td>15</td>
<td>495</td>
<td>0.38</td>
<td>1.42</td>
</tr>
<tr>
<td>P14869</td>
<td>60 S acidic ribosomal protein P0</td>
<td>34</td>
<td>34,336</td>
<td>102</td>
<td>9</td>
<td>379</td>
<td>0.50</td>
<td>2.16</td>
</tr>
<tr>
<td>P03975</td>
<td>IgE-binding protein</td>
<td>35</td>
<td>63,221</td>
<td>122</td>
<td>10</td>
<td>368</td>
<td>0.33</td>
<td>1.15</td>
</tr>
<tr>
<td>Q9C203</td>
<td>Glycyl-ribosomal isocitrate synthase</td>
<td>37</td>
<td>82,624</td>
<td>77</td>
<td>9</td>
<td>341</td>
<td>0.21</td>
<td>0.62</td>
</tr>
<tr>
<td>P35215</td>
<td>14-3-3 protein ε β</td>
<td>40</td>
<td>27,925</td>
<td>381</td>
<td>12</td>
<td>401</td>
<td>0.75</td>
<td>4.62</td>
</tr>
<tr>
<td>P42932</td>
<td>T-complex protein 1, α subunit</td>
<td>42</td>
<td>60,088</td>
<td>96</td>
<td>9</td>
<td>281</td>
<td>0.26</td>
<td>0.81</td>
</tr>
<tr>
<td>P51881</td>
<td>ADP, ATP carrier protein, fibroblast isoform</td>
<td>46</td>
<td>33,136</td>
<td>102</td>
<td>8</td>
<td>294</td>
<td>0.42</td>
<td>1.64</td>
</tr>
<tr>
<td>Q0JIK5</td>
<td>Nucleolar RNA helicase II</td>
<td>48</td>
<td>94,151</td>
<td>30</td>
<td>16</td>
<td>268</td>
<td>0.16</td>
<td>0.45</td>
</tr>
<tr>
<td>P14148</td>
<td>60 S ribosomal protein L7</td>
<td>52</td>
<td>31,457</td>
<td>120</td>
<td>8</td>
<td>227</td>
<td>0.53</td>
<td>2.38</td>
</tr>
<tr>
<td>Q9WAA4</td>
<td>Transgelin</td>
<td>65</td>
<td>23,810</td>
<td>130</td>
<td>8</td>
<td>258</td>
<td>0.62</td>
<td>3.12</td>
</tr>
<tr>
<td>P14211</td>
<td>Calreticulin precursor</td>
<td>72</td>
<td>48,136</td>
<td>114</td>
<td>8</td>
<td>246</td>
<td>0.33</td>
<td>1.15</td>
</tr>
<tr>
<td>P16858</td>
<td>Glyceraldehyde-3-phosphate dehydrogenase</td>
<td>87</td>
<td>59,541</td>
<td>400</td>
<td>8</td>
<td>270</td>
<td>0.53</td>
<td>2.41</td>
</tr>
<tr>
<td>P29314</td>
<td>40 S ribosomal protein S9</td>
<td>88</td>
<td>22,418</td>
<td>135</td>
<td>6</td>
<td>201</td>
<td>0.46</td>
<td>1.89</td>
</tr>
<tr>
<td>Q60932</td>
<td>Voltage-dependent anion-selective channel protein</td>
<td>89</td>
<td>32,502</td>
<td>72</td>
<td>6</td>
<td>261</td>
<td>0.38</td>
<td>1.37</td>
</tr>
<tr>
<td>P17080</td>
<td>GTP-binding nuclear protein RAN</td>
<td>97</td>
<td>24,579</td>
<td>255</td>
<td>5</td>
<td>181</td>
<td>0.45</td>
<td>1.85</td>
</tr>
<tr>
<td>P17008</td>
<td>40 S ribosomal protein S16</td>
<td>98</td>
<td>16,418</td>
<td>456</td>
<td>6</td>
<td>174</td>
<td>0.60</td>
<td>2.98</td>
</tr>
<tr>
<td>Q60930</td>
<td>Voltage-dependent anion-selective channel protein</td>
<td>99</td>
<td>32,340</td>
<td>54</td>
<td>4</td>
<td>161</td>
<td>0.27</td>
<td>0.85</td>
</tr>
<tr>
<td>P11983</td>
<td>T-complex protein 1, α subunit B</td>
<td>100</td>
<td>60,867</td>
<td>36</td>
<td>6</td>
<td>143</td>
<td>0.18</td>
<td>0.52</td>
</tr>
<tr>
<td>P05064</td>
<td>Fruuctose-bisphosphate aldolase A</td>
<td>103</td>
<td>39,656</td>
<td>210</td>
<td>6</td>
<td>260</td>
<td>0.33</td>
<td>1.15</td>
</tr>
<tr>
<td>P09058</td>
<td>40 S ribosomal protein S8</td>
<td>109</td>
<td>24,344</td>
<td>96</td>
<td>6</td>
<td>186</td>
<td>0.60</td>
<td>2.98</td>
</tr>
<tr>
<td>Q01320</td>
<td>DNA topoisomerase II, α isozyme</td>
<td>143</td>
<td>173,567</td>
<td>36</td>
<td>4</td>
<td>125</td>
<td>0.05</td>
<td>0.12</td>
</tr>
<tr>
<td>Q8VEW8</td>
<td>Phosphatase carrier protein, mitochondrial precursor</td>
<td>149</td>
<td>40,065</td>
<td>48</td>
<td>4</td>
<td>122</td>
<td>0.19</td>
<td>0.55</td>
</tr>
<tr>
<td>P19253</td>
<td>60 S ribosomal protein L13a</td>
<td>150</td>
<td>23,432</td>
<td>48</td>
<td>4</td>
<td>123</td>
<td>0.33</td>
<td>1.15</td>
</tr>
<tr>
<td>P08526</td>
<td>60 S ribosomal protein L27</td>
<td>157</td>
<td>15,657</td>
<td>86</td>
<td>3</td>
<td>113</td>
<td>0.50</td>
<td>2.16</td>
</tr>
<tr>
<td>P47961</td>
<td>40 S ribosomal protein S4</td>
<td>160</td>
<td>29,666</td>
<td>162</td>
<td>4</td>
<td>109</td>
<td>0.21</td>
<td>0.62</td>
</tr>
<tr>
<td>Q06647</td>
<td>ATP synthase oligomycin sensitivity conferral protein</td>
<td>179</td>
<td>23,440</td>
<td>78</td>
<td>3</td>
<td>98</td>
<td>0.23</td>
<td>0.70</td>
</tr>
<tr>
<td>Q9CPDR</td>
<td>60 S ribosomal protein L17</td>
<td>182</td>
<td>21,506</td>
<td>90</td>
<td>3</td>
<td>96</td>
<td>0.30</td>
<td>1.00</td>
</tr>
<tr>
<td>P39026</td>
<td>60 S ribosomal protein L11</td>
<td>186</td>
<td>20,337</td>
<td>108</td>
<td>3</td>
<td>92</td>
<td>0.33</td>
<td>1.15</td>
</tr>
<tr>
<td>Q091F9</td>
<td>60 S ribosomal protein L34</td>
<td>204</td>
<td>13,381</td>
<td>96</td>
<td>3</td>
<td>83</td>
<td>0.50</td>
<td>2.16</td>
</tr>
<tr>
<td>P08807</td>
<td>Peroxidoxin 1</td>
<td>206</td>
<td>31,261</td>
<td>72</td>
<td>4</td>
<td>83</td>
<td>0.27</td>
<td>0.85</td>
</tr>
<tr>
<td>Q62188</td>
<td>Dihydropyrimidinase related protein-3</td>
<td>207</td>
<td>62,296</td>
<td>30</td>
<td>3</td>
<td>82</td>
<td>0.10</td>
<td>0.27</td>
</tr>
<tr>
<td>P50310</td>
<td>Phosphoglycerate kinase</td>
<td>213</td>
<td>44,776</td>
<td>30</td>
<td>3</td>
<td>80</td>
<td>0.13</td>
<td>0.33</td>
</tr>
<tr>
<td>Q9DBJ1</td>
<td>Phosphoglycerate mutase 1</td>
<td>223</td>
<td>28,797</td>
<td>114</td>
<td>3</td>
<td>70</td>
<td>0.25</td>
<td>0.78</td>
</tr>
<tr>
<td>P11442</td>
<td>Calreticulin heavy chain</td>
<td>226</td>
<td>193,187</td>
<td>55</td>
<td>3</td>
<td>69</td>
<td>0.04</td>
<td>0.10</td>
</tr>
</tbody>
</table>

* Protein concentrations were measured by “reversed” isotope dilution using SILAC-labeled proteins and unlabeled synthetic peptides.
pared with QSTAR did not provide better correlation between emPAI and protein abundance. This could be because of the limited capacity of LTQ to trap ions even in the linear configuration. We also evaluated the influence of sample complexity by using multidimensional chromatography (23, 31). As shown in Fig. 2E, SCX fractionation gave improvement in emPAI accuracy. To confirm the effect of the reduction of sample complexity on emPAI accuracy, we used both QSTAR and LTQ in combination with SCX fractionation and obtained in total 2752 identified proteins with 11,727 non-redundant peptides from neuro2a cells. The correlation between emPAI values of the 46 test proteins and their protein abundances was significantly improved as shown in Fig. 2F. Note that PAI values of 22 proteins of 46 proteins were more than one in this analysis, whereas only two proteins had PAI values of more than one in the QSTAR analysis without SCX fractionation. This result shows that under the current conditions emPAI was not saturated. However, it would be possible to saturate emPAI if some proteins are highly abundant. We observed this in our analysis of the malaria proteome where hemoglobin was extremely abundant because the samples were prepared from red blood cells (15). Extremely abundant proteins may furthermore affect the efficiency of protein identification because of ionization suppression and detector saturation as well as the limited loading capacity of LC columns. The removal of extremely abundant proteins is therefore required to improve the identification efficiency and can be achieved by gel-enhanced LC-MS (one-dimensional gel followed by slicing, digesting, and LC-MS analysis) as shown in our malaria proteome study or albumin depletion treatment for plasma proteome studies. Such a treatment will also remove the influence of emPAI saturation.

We also examined the influence of the injected sample amounts on the emPAI-based molar fractions. Using the whole cell lysate of neuro2a cells, three different levels (basal and 3× and 9× dilutions) were analyzed by LC-MS/MS. For 20 proteins with commonly identified peptides in all three analyses, constant values of the molar fraction were obtained (deviation factors were 1.66 ± 0.55 for 3× dilution and 1.85 ± 0.85 for 9× dilution, respectively), whereas emPAI values depended on the injected amounts as expected.

**Application to Comprehensive Protein Expression Analysis**—The emPAI is a convenient and easily obtained index that can be used to produce protein expression data from any LC-MS/MS runs. We applied this approach to obtain data for the relative expression levels in HCT116 cells. Experimental details are described under “Materials and Methods.”

**Fig. 3. Comparison between gene and protein expression levels in HCT116 cells.**
comparison with gene expression data in HCT116 human cancer cells. A DNA microarray provided expression data for 4971 genes, whereas a single LC-MS run provided 402 identified proteins based on 1811 peptides with unique sequences. Bridging gene symbols with protein accession numbers resulted in a total of 227 genes/protein pairs for the expression comparison study. A weak correlation was observed in Fig. 3 as expected from previous studies on yeast (19, 32). Interestingly most of the outliers were ribosomal proteins. It is well known that, unlike prokaryotes such as *Escherichia coli*, mammalian cells regulate the expression levels of ribosomal proteins not only by transcription but also at the steps of transport of mRNA and translation and by degradation of excess amounts of proteins not associated with rRNA (33, 34). Accordingly in a comparison study between gene and protein expression levels using emPAI for *E. coli*, we did not find such a deviation of ribosomal proteins. Although both gene and protein expression data are not sufficiently accurate to discriminate a 10% difference, for instance, it is quite helpful to obtain a broad overview as shown above. We also note that the protein quantitation error of our simple emPAI is similar or better than the error in determining mRNA expression in DNA microarrays.

Conclusions—We have established a scale for estimating absolute protein abundance named emPAI. Because emPAI is easily calculated from the output information of database search engines such as Mascot, it is possible to apply this approach to previously measured or published datasets to add quantitative information without any additional steps. emPAI can also be used for relative quantitation especially in cases where isotope-based approaches cannot be applied because of quantitative changes that are too large for accurate measurements of ratios, because metabolic labeling is not possible, or because sensitivity constraints do not allow chemical labeling techniques. In such cases, emPAI values of proteins in one sample can compare with those in another sample, and the outliers from the emPAI correlation between two samples can be determined as increasing or decreasing proteins.

This emPAI approach was applied to multidimensional separation-MS/MS to extend the coverage of proteins. Further improvement would be possible by optimizing MS instrument-dependent parameters such as ionization dependence on m/z region. Because the emPAI index can be calculated with a simple script and does not require further experimentation in protein identification experiments, we suggest its routine use in the reporting of proteomic results.

Acknowledgments—We thank Rie Ushijima, Junro Kuromitsu, Takashi Owa, and Akira Yokoi (LSFT, Eisai) for DNA microarray experiments and Norimasa Miyamoto (LSFT) for SILAC cell culture. We also thank Peter Mortensen and Shao-En Ong (CEBI) for MSQuant setting and members of LSFT and CEBI for fruitful discussion. Y. I. thanks Eisai for the opportunity to stay in CEBI.

* Work at LSFT and CEBI was supported by NEDO (New Energy and Industrial Technology Development Organization, Japan) and the Danish National Research Foundation, respectively. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

** To whom correspondence may be addressed. Tel.: 81-29-847-7192; Fax: 81-29-847-7614; E-mail: y-ishihama@hhc.eisai.co.jp.

REFERENCES


* Y. Ishihama, D. Frishman, and M. Mann, unpublished data.


