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

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 10^{PAI} 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 quantitation 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

From the ‡Laboratory of Seeds Finding Technology (LSFT), Eisai Co., Ltd., 5-1-3 Tokodai, Tsukuba, Ibaraki 300-2635, Japan, the §Center for Experimental Bioinformatics (CEBI), University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark, and ‖The FIRC Institute for Molecular Oncology, 20139 Milan, Italy

Received, March 3, 2005, and in revised form, June 14, 2005

Published, MCP Papers in Press, June 14, 2005, DOI 10.1074/mcp.M500061-MCP200

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 M 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 (Nikkyo 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_{18} 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-fluorenyl)methoxycarbonyl) 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.

Nano-LC-MS/MS Analysis—All samples were analyzed by nano-LC-MS/MS using a QSTAR Pulsar i (AB/MDS-Sciex, Toronto, Canada), a Finnigan LCQ 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-μm ports. ReproSil C_{18} materials (3 μm, Dr. Maisch, Ammerbuch, Germany) were packed into a self-pulled needle (100-μm inner diameter, 6-μm opening, 150-mm length) with a nitrogen-pressurized column loader cell (Nikkyo) to prepare an analytical column needle with "stone-arch" frit (24). A Teflon-coated column holder (Nikkyo) 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 μl, 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/bitt/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

$$\mathrm{PAI} = \frac{N_{\mathrm{obsd}}}{N_{\mathrm{obsbl}}} \qquad \text{(Eq. 1)}$$

where N_{obsd} and N_{obsbl} are the number of observed peptides per

[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.

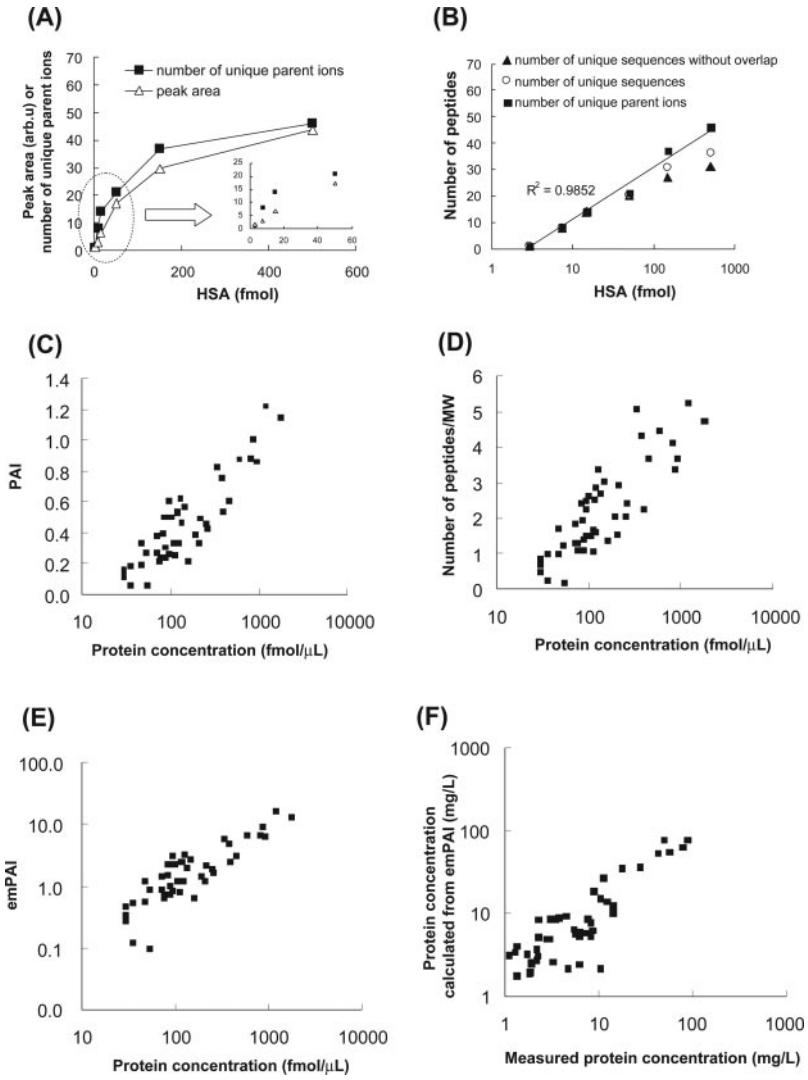

FIG. 1. **Relationship between protein concentration and several parameters.** *A*, peak area and the number of unique parent ions of peptides *versus* injection amount of HSA. The most abundant tryptic peptide of HSA, LCTVATLR, was used for peak area measurement. *B*, numbers of peptides counted in three different ways *versus* injection amount of HSA. *C*, protein concentration *versus* PAI for 46 proteins in neuro2a cells. *D*, protein concentration *versus* the number of peptides divided by molecular weight of proteins for 46 proteins in neuro2a cells. *E*, relationship between protein concentration and emPAI for 46 proteins in neuro2a cells. *F*, absolute quantitation of 46 proteins in neuro2a cells using emPAI. QSTAR with faster scans (0.55 s for each MS/MS scan) was used for these experiments. Protein concentrations in neuro2a cells were measured by spiking synthetic peptides to neuro2a cells as described under "Materials and Methods." *arb.u*, arbitrary units.

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

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

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

$$\text{Protein content (mol \%)} = \frac{\text{emPAI}}{\Sigma\,(\text{emPAI})} \times 100 \qquad \text{(Eq. 3)}$$

$$\text{Protein content (weight \%)} = \frac{\text{emPAI} \times M_r}{\Sigma\,(\text{emPAI} \times M_r)} \times 100 \qquad \text{(Eq. 4)}$$

where M_r is the molecular weight of the protein, and Σ(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

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

where measured values are larger than estimated values or

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

where estimated values are larger than measured values.

DNA Microarray Analysis—HCT116-C9 cells were plated at 5.0×10^6 cells/dish in 10-cm-diameter dishes with 10 ml of the culture medium. After 24-h preincubation, the cells were treated for 12 h with 0.015% DMSO. Duplicate experiments were performed using Affymetrix HuGene FL arrays according to established protocols. Affymetrix GeneChip software was used to extract gene signal intensities, and two sets of data were grouped and averaged based on gene symbols.

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. 1*A*, both peak area and the number of identified peptides with unique parent ions increased as the injection amount increased, although both curves saturated at

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. 1*B*). 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. 1*B* shows that the number of peptides based on unique parent ions (Fig. 1*B*, *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 proteins were identified based on 1462 peptides. For accurate absolute quantitation, we spiked 54 synthetic peptides containing [$^{12}C_6$]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. 1*C* 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. 1*C*, $r = 0.89$, deviation factor (average $\pm$ S.D.) $= 1.6 \pm 0.5$) followed by number of peptides divided by protein M_r (Fig. 1*D*, $r = 0.84$, deviation factor $= 1.8 \pm 0.8$), 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 $= 2.7 \pm 2.4$ and $r = 0.71$, deviation factor $= 2.7 \pm 2.6$, 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. 1*E*. 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. 1*F*, 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

TABLE I
46 proteins identified and quantified in mouse neuro2a cells

Swis-Prot accession no.	Protein name	Mascot hit no.	M_r	Protein concentration[a]	Number of observed peptides	Mascot score	PAI	emPAI
				fmol/μl				
P19378	Heat shock cognate 71-kDa protein	1	70,989	822	29	1235	0.88	6.56
P07901	Heat shock protein HSP 90-α	2	85,003	940	31	1047	0.86	6.26
P20152	Vimentin	5	53,581	336	27	1060	0.82	5.58
P58252	Elongation factor 2 (EF-2)	6	96,091	118	24	830	0.53	2.41
Q03265	ATP synthase α chain, mitochondrial precursor	8	59,830	149	18	635	0.56	2.65
P17182	α enolase	9	47,322	596	21	828	0.88	6.50
P15331	Peripherin	10	54,349	84	13	556	0.39	1.48
P48975	Actin, cytoplasmic 1 (β-actin)	12	42,053	1206	22	894	1.22	15.68
P05213	Tubulin α-2 chain (α-tubulin 2)	14	50,818	1782	24	862	1.14	12.89
P52480	Pyruvate kinase, M2 isozyme	16	58,289	216	17	643	0.49	2.06
P20001	Elongation factor 1-α 1	24	50,424	870	17	647	1.00	9.00
P08113	Endoplasmin precursor	25	92,703	90	10	379	0.23	0.71
O35501	Stress-70 protein, mitochondrial precursor	27	73,970	195	15	495	0.38	1.42
P14869	60 S acidic ribosomal protein P0	34	34,336	102	9	379	0.50	2.16
P03975	IgE-binding protein	35	63,221	122	10	368	0.33	1.15
Q9CZD3	Glycyl-tRNA synthetase	37	82,624	77	9	341	0.21	0.62
P35215	14-3-3 protein ζ/δ	40	27,925	381	12	401	0.75	4.62
P42932	T-complex protein 1, θ subunit	42	60,088	96	9	281	0.26	0.81
P51881	ADP, ATP carrier protein, fibroblast isoform	46	33,138	264	8	294	0.42	1.64
Q9JIK5	Nucleolar RNA helicase II	48	94,151	30	8	268	0.16	0.45
P14148	60 S ribosomal protein L7	52	31,457	120	9	227	0.53	2.38
Q9WVA4	Transgelin 2	65	23,810	130	8	258	0.62	3.12
P14211	Calreticulin precursor	72	48,136	114	8	246	0.33	1.15
P16858	Glyceraldehyde-3-phosphate dehydrogenase	87	35,941	400	8	270	0.53	2.41
P29314	40 S ribosomal protein S9	88	22,418	135	6	201	0.46	1.89
Q60932	Voltage-dependent anion-selective channel protein	89	32,502	72	6	261	0.38	1.37
P17080	GTP-binding nuclear protein RAN	97	24,579	255	5	181	0.45	1.85
P17008	40 S ribosomal protein S16	98	16,418	456	6	174	0.60	2.98
Q60930	Voltage-dependent anion-selective channel protein	99	32,340	54	4	161	0.27	0.85
P11983	T-complex protein 1, α subunit B	100	60,867	36	6	143	0.18	0.52
P05064	Fructose-bisphosphate aldolase A	103	39,656	210	6	260	0.33	1.15
P09058	40 S ribosomal protein S8	109	24,344	96	6	186	0.60	2.98
Q01320	DNA topoisomerase II, α isozyme	143	173,567	36	4	125	0.05	0.12
Q8VEM8	Phosphate carrier protein, mitochondrial precursor	149	40,063	48	4	122	0.19	0.55
P19253	60 S ribosomal protein L13a	150	23,432	48	4	123	0.33	1.15
P08526	60 S ribosomal protein L27	157	15,657	86	3	113	0.50	2.16
P47961	40 S ribosomal protein S4	160	29,666	162	4	109	0.21	0.62
Q06647	ATP synthase oligomycin sensitivity conferral protein	179	23,440	78	3	98	0.23	0.70
Q9CPR4	60 S ribosomal protein L17	182	21,506	90	3	96	0.30	1.00
P39026	60 S ribosomal protein L11	186	20,337	108	3	92	0.33	1.15
Q9D1R9	60 S ribosomal protein L34	204	13,381	96	3	83	0.50	2.16
O08807	Peroxiredoxin 4	206	31,261	72	4	83	0.27	0.85
Q62188	Dihydropyrimidinase related protein-3	207	62,296	30	3	82	0.10	0.27
P50310	Phosphoglycerate kinase	213	44,776	30	3	80	0.13	0.33
Q9DBJ1	Phosphoglycerate mutase 1	223	28,797	114	3	70	0.25	0.78
P11442	Clathrin heavy chain	226	193,187	55	3	69	0.04	0.10

[a] Protein concentrations were measured by "reversed" isotope dilution using SILAC-labeled proteins and unlabeled synthetic peptides.

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. 2*C*) 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. 2*D*). However, the faster scan cycle of LTQ com-

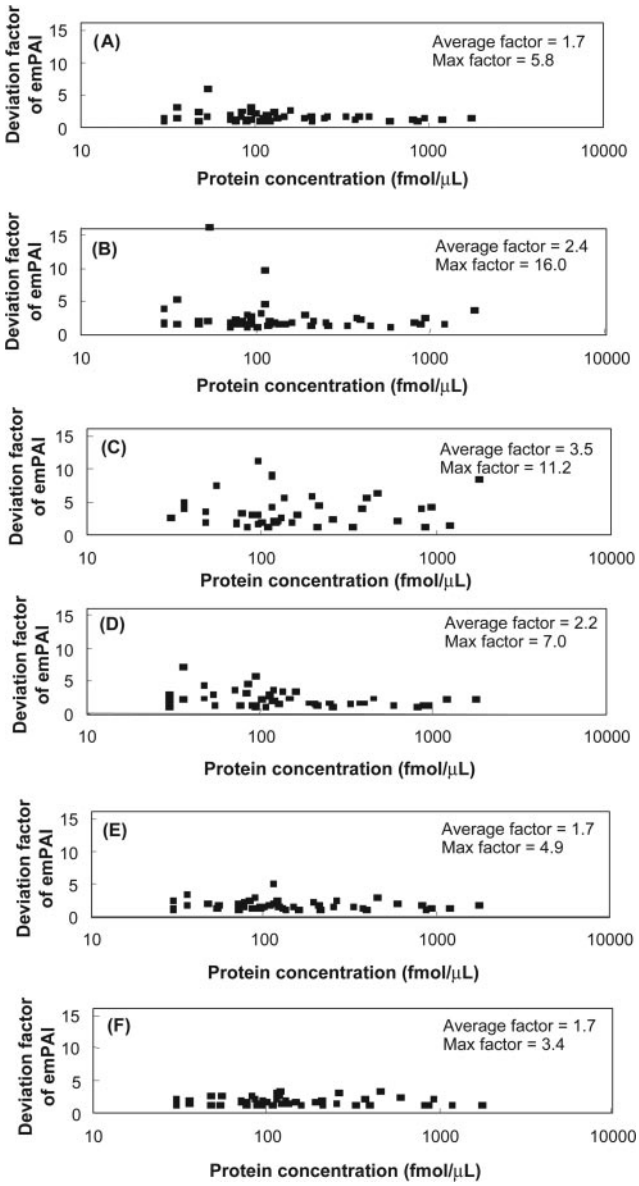

FIG. 2. **Influence of MS measurement conditions on the deviation factors between the estimated and measured concentrations of 46 proteins in neuro2a cells.** *A*, QSTAR with faster scan cycle (0.55 s for each MS/MS scan). *B*, QSTAR with slower scan cycle (1.5 s for each MS/MS scan). *C*, LCQ with slower scan cycle (1.2 s on average for each MS/MS scan). *D*, LTQ with faster scan cycle (0.38 s on average for each MS/MS scan). *E*, multidimensional LC-MS/MS with LTQ (five fractions). *F*, multidimensional LC-MS/MS with QSTAR and LTQ.

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. 2*E*, 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

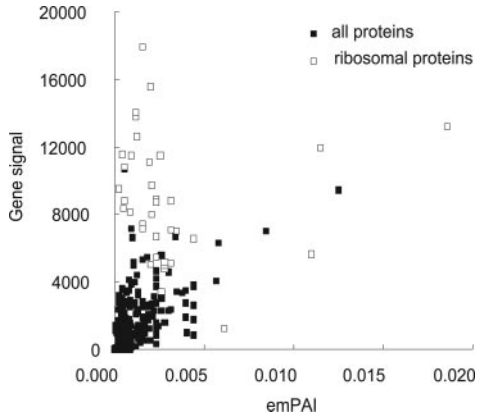

FIG. 3. **Comparison between gene and protein expression levels in HCT116 cells.** Experimental details are described under "Materials and Methods."

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. 2*F*. 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

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.[2] 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.

[2] Y. Ishihama, D. Frishman, and M. Mann, unpublished data.

* 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.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

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

** To whom correspondence may be addressed. Tel.: 45-6550-2364; Fax: 45-6593-3929; E-mail: mann@bmb.sdu.dk.

REFERENCES

1. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
2. Oda, Y., Huang, K., Cross, F. R., Cowburn, D., and Chait, B. T. (1999) Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 6591–6596
3. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999
4. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1,** 376–386
5. MacCoss, M. J., Wu, C. C., Liu, H., Sadygov, R., and Yates, J. R., III (2003) A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal. Chem.* **75,** 6912–6921
6. Foster, L. J., De Hoog, C. L., and Mann, M. (2003) Unbiased quantitative proteomics of lipid rafts reveals high specificity for signaling factors. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 5813–5818
7. Blagoev, B., Kratchmarova, I., Ong, S. E., Nielsen, M., Foster, L. J., and Mann, M. (2003) A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nat. Biotechnol.* **21,** 315–318
8. Ranish, J. A., Yi, E. C., Leslie, D. M., Purvine, S. O., Goodlett, D. R., Eng, J., and Aebersold, R. (2003) The study of macromolecular complexes by quantitative proteomics. *Nat. Genet.* **33,** 349–355
9. Schulze, W. X., and Mann, M. (2004) A novel proteomic screen for peptide-protein interactions. *J. Biol. Chem.* **279,** 10756–10764
10. Oda, Y., Owa, T., Sato, T., Boucher, B., Daniels, S., Yamanaka, H., Shinohara, Y., Yokoi, A., Kuromitsu, J., and Nagasu, T. (2003) Quantitative chemical proteomics for identifying candidate drug targets. *Anal. Chem.* **75,** 2159–2165
11. Barr, J. R., Maggio, V. L., Patterson, D. G., Jr., Cooper, G. R., Henderson, L. O., Turner, W. E., Smith, S. J., Hannon, W. H., Needham, L. L., and Sampson, E. J. (1996) Isotope dilution–mass spectrometric quantification of specific proteins: model application with apolipoprotein A-I. *Clin. Chem.* **42,** 1676–1682
12. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., and Gygi, S. P. (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 6940–6945
13. Havlis, J., and Shevchenko, A. (2004) Absolute quantification of proteins in solutions and in polyacrylamide gels by mass spectrometry. *Anal. Chem.* **76,** 3029–3036
14. Corbin, R. W., Paliy, O., Yang, F., Shabanowitz, J., Platt, M., Lyons, C. E., Jr., Root, K., McAuliffe, J., Jordan, M. I., Kustu, S., Soupene, E., and Hunt, D. F. (2003) Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 9232–9237
15. Lasonder, E., Ishihama, Y., Andersen, J. S., Vermunt, A. M., Pain, A., Sauerwein, R. W., Eling, W. M., Hall, N., Waters, A. P., Stunnenberg, H. G., and Mann, M. (2002) Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. *Nature* **419,** 537–542
16. Shen, Y., Zhao, R., Berger, S. J., Anderson, G. A., Rodriguez, N., and Smith, R. D. (2002) High-efficiency nanoscale liquid chromatography coupled on-line with mass spectrometry using nanoelectrospray ionization for proteomics. *Anal. Chem.* **74,** 4235–4249

17. Rappsilber, J., Ryder, U., Lamond, A. I., and Mann, M. (2002) Large-scale proteomic analysis of the human spliceosome. *Genome. Res.* **12,** 1231–1245
18. Sanders, S. L., Jennings, J., Canutescu, A., Link, A. J., and Weil, P. A. (2002) Proteomics of the eukaryotic transcription machinery: identification of proteins associated with components of yeast TFIID by multidimensional mass spectrometry. *Mol. Cell. Biol.* **22,** 4723–4738
19. Liu, H., Sadygov, R. G., and Yates, J. R., III (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76,** 4193–4201
20. Cox, B. J., Kislinger, T., Wigle, D. A., Brown, K., Manning, D., Jurisica, I., Emili, A., and Rossant, J. (2004) in *Proceedings of the 52nd ASMS Conference on Mass Spectrometry and Allied Topics, May 23–27, 2004, Nashville*, Abstr. ThPS352, American Society for Mass Spectrometry, Santa Fe, NM
21. Allet, N., Barrillat, N., Baussant, T., Boiteau, C., Botti, P., Bougueleret, L., Budin, N., Canet, D., Carraud, S., Chiappe, D., Christmann, N., Colinge, J., Cusin, I., Dafflon, N., Depresle, B., Fasso, I., Frauchiger, P., Gaertner, H., Gleizes, A., Gonzalez-Couto, E., Jeandenans, C., Karmime, A., Kowall, T., Lagache, S., Mahe, E., Masselot, A., Mattou, H., Moniatte, M., Niknejad, A., Paolini, M., Perret, F., Pinaud, N., Ranno, F., Raimondi, S., Reffas, S., Regamey, P. O., Rey, P. A., Rodriguez-Tome, P., Rose, K., Rossellat, G., Saudrais, C., Schmidt, C., Villain, M., and Zwahlen, C. (2004) *In vitro* and *in silico* processes to identify differentially expressed proteins. *Proteomics* **4,** 2333–2351
22. Rappsilber, J., Ishihama, Y., and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75,** 663–670
23. Ishihama, Y., Sato, T., Tabata, T., Miyamoto, N., Sagane, K., Nagasu, T., and Oda, Y. (2005) Quantitative mouse brain proteomics using culture-derived isotope tags as internal standards. *Nat. Biotechnol.* **23,** 617–621
24. Ishihama, Y., Rappsilber, J., Andersen, J. S., and Mann, M. (2002) Microcolumns with self-assembled particle frits for proteomics. *J. Chromatogr. A* **979,** 233–239
25. Meek, J. L. (1980) Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. *Proc. Natl. Acad. Sci. U. S. A.* **77,** 1632–1636
26. Sakamoto, Y., Kawakami, N., and Sasagawa, T. (1988) Prediction of peptide retention times. *J. Chromatogr.* **442,** 69–79
27. Sweetman, G., Bantscheff, M., Schirle, M., Rick, J., and Kuster, B. (2004) in *Proceedings of the 52nd ASMS Conference on Mass Spectrometry and Allied Topics, May 23–27, 2004, Nashville*, Abstr. TPR361, American Society for Mass Spectrometry, Santa Fe, NM
28. Read, S. M., and Northcote, D. H. (1981) Minimization of variation in the response to different proteins of the Coomassie blue G dye-binding assay for protein. *Anal. Biochem.* **116,** 53–64
29. Hager, J. W. (2002) A new linear ion trap mass spectrometer. *Rapid Commun. Mass Spectrom.* **16,** 512–526
30. Schwartz, J. C., Senko, M. W., and Syka, J. E. (2002) A two-dimensional quadrupole ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **13,** 659–669
31. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R., III (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17,** 676–682
32. Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19,** 1720–1730
33. Tsurugi, K. (1989) On the regulation of ribosome synthesis in eukaryotic cell. *Seikagaku* **61,** 271–284
34. Mager, W. H. (1988) Control of ribosomal protein gene expression. *Biochim. Biophys. Acta* **949,** 1–15