BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 4 2010, pages 576–577
doi:10.1093/bioinformatics/btp700

Gene expression

emPAI Calc—for the estimation of protein abundance from large-scale identification data by liquid chromatography-tandem mass spectrometry

Kosaku Shinoda[1,2], Masaru Tomita[1,2] and Yasushi Ishihama[2,3,*]

[1]Human Metabolome Technologies, Inc., Tsuruoka, Yamagata, 997-0052, [2]Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata, 997–0017 and [3]PRESTO, Japan Science and Technology Agency, Sanbancho bldg., 5-Sanbancho, Chiyodaku, Tokyo 102-0075, Japan

Received on June 16, 2009; revised on November 27, 2009; accepted on December 17, 2009

Advance Access publication December 22, 2009

Associate Editor: David Rocke

ABSTRACT

Summary: emPAI Calc is an open-source web application for the estimation of protein abundance. It uses the correlation between the number of identified peptides and protein abundance in mass spectrometry-based proteomic experiments. The program is the first implementation of our previously reported emPAI algorithm; it calculates the emPAI from the protein identification results obtained by database search engines such as Mascot.™

Availability: http://empai.iab.keio.ac.jp/; http://empai.iab.keio.ac.jp/supplement.php Source codes are available under Mozilla Public License.

Contact: y-ishi@ttck.keio.ac.jp

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Proteomic liquid chromatography (LC)-tandem mass spectrometry (MS/MS) approaches combined with genome-annotated databases currently allow the identification of thousands of proteins from complex mixtures. However, a comprehensive approach for the absolute quantification of proteins has yet to be established. Protein concentrations are fundamental parameters because the kinetics/dynamics of the cellular machinery are described in terms of changes in the concentration of proteins in particular compartments. *In vivo* quantitative information has been derived from the intensity of gel staining, however, in complex-mixture analysis, proteins cannot be stained individually and most information about protein abundance is lost.

Even a single LC-MS/MS analysis can generate a long list of identified proteins with the help of database searching and additional information (such as the hit rank in identification, the probability score, the number of identified peptides and the ion counts of identified peptides and the LC retention times) can be extracted. Qualitatively, some parameters such as the hit rank and the score and number of peptides per protein (Corbin *et al*., 2003) can be considered as indicators of protein abundance in an analyzed sample. Among these, the integrated ion counts of the peptides identifying each protein represent the most direct parameter to describe abundance (Lasonder *et al*., 2002). However, MS is not as versatile as absorbance detection because of its limited linearity and ionization–suppression effects (Shen *et al*., 2002). Therefore, to obtain at least approximate quantitative information, these parameters must be normalized.

To our knowledge, the first quantitative approach to achieve this goal was based on the number of observed peptides per protein normalized by the theoretical number of peptides, the so-called protein abundance index (PAI) (Rappsilber *et al*., 2002). More recently, several groups reported new identification-based algorithms for protein-abundance estimation accompanied by large-scale validation results including the spectral count (SC) and its derivatives (Liu *et al*., 2004; Lu *et al*., 2007; Zybailov *et al*., 2006), and the exponentially modified PAI (emPAI) (Ishihama *et al*., 2005).

EmPAI showed a high ($r = 0.89$) correlation with the actual protein amount in complex mixtures of mouse neuro2A cells with a wide dynamic range from 30 fmol to 1.8 pmol/μl in the sample solution (Ishihama *et al*., 2005). Although the emPAI is not as accurate as quantification using synthesized peptide standards, it is quite useful for obtaining a broad overview of proteome profiles. However, wet bench biologists not familiar with proteomic experiments may find it difficult to calculate the emPAI.

We present emPAI Calc, a web application for the calculation of the emPAI from LC-MS/MS data. The number of observed peptides is extracted from the result files of Mascot™ (Matrix Sciences, London, UK) search engines with a user-configurable identification threshold. The theoretical number of peptides (observable peptides) is aggregated from appropriate protein databanks and the peptides are filtered using their molecular weight (MW) and the predicted LC retention times. As emPAI Calc was developed to correspond to multiple LC conditions it is applicable to most LC-MS-based proteomic experiments. In addition, it can integrate multiple LC-MS/MS results into one emPAI calculation. This feature is useful for large-scale proteomic experiments that require sample pre-fractionation, such as protein-level gel electrophoresis and/or peptide-level ion-exchange chromatography prior to LC-MS/MS analysis. Furthermore, emPAI Calc can be applied to previously measured or published datasets to add quantitative information without additional steps. As a typical example and to demonstrate the

*To whom correspondence should be addressed.

accuracy of emPAI, the correlation between the emPAI calculated from published yeast proteome data (de Godoy *et al*., 2006) and protein abundance based on quantitative Western blotting is shown in Supplementary Figure 1.

2 DESCRIPTION OF THE ALGORITHM

The MS signal for any given peptide is determined by several factors, the most important of which is its ionizability in electrospray; therefore, it is difficult to directly quantify proteins in an LC-MS/MS experiment. However, there is a general correlation between the number of peptides sequenced per protein and the amount of protein present in the mixture. As larger proteins can give rise to more peptides, Rappsilber *et al*. (2002) employed the PAI, which represents the number of peptides identified divided by the number of theoretically observable tryptic peptides. Although the PAI can estimate the abundance relationship between proteins, it cannot express the molar fraction directly. Interestingly, the number of peptides exhibits a linear relationship with the logarithm of the quantity of injected proteins. Therefore, we converted the PAI to the emPAI (that is, $10^{\mathrm{PAI}}-1$), and showed that the former is directly proportional to the protein content (Ishihama *et al*., 2005).

To calculate the number of observable peptides per protein, the proteins were digested *in silico* and the mass of the obtained peptides was compared with the MS scan range. In addition, the expected retention times under LC conditions were predicted using multiple linear regressions (Meek, 1980) or artificial neural networks (Petritis *et al*., 2003). *In silico* peptides that fell outside of the MW/retention time range were eliminated. Detailed procedures for the calculation of observable peptides are presented in Supplementary Text. With respect to the number of observed peptides we adopted counting of unique parent ions, including different charge states from the same peptide sequence. We noted that for some proteins the number of observed peptides is greater than the number of observable peptides. The PAI is calculated as follows:

$$\mathrm{PAI}=\frac{N_{\mathrm{obst}}}{N_{\mathrm{obsbl}}} \tag{1}$$

where, N_{obst} and N_{obsbl} are the number of observed unique parent ions per protein and the number of observable peptides per protein, respectively . The emPAI is defined as follows:

$$\mathrm{emPAI}=10^{\mathrm{PAI}}-1 \tag{2}$$

Thus, the protein contents in the molar fraction percentage is described by:

$$\text{Protein content (mol \%)}=\frac{\mathrm{emPAI}}{\sum(\mathrm{emPAI})}\times 100 \tag{3}$$

Here, $\Sigma_{(\mathrm{emPAI})}$ is the summation of the emPAI values for all of the identified proteins. The detailed procedure for the emPAI calculation is described in Supplementary Figure 2 and in a previous paper (Ishihama *et al*., 2005).

3 IMPLEMENTATION

emPAI Calc is accessible from an intuitive web interface. It provides automated calculations from Mascot™ search results or CSV format files, to which the result files from other search engines (such as Sequest) can be converted. For file uploads, Uber-uploader (http://uber-uploader.sourceforge.net), distributed under the Mozilla Public Licence, is used as part of emPAI Calc. Queried files are parsed using Perl and the number of observed peptides, the identification scores, protein identifications (IDs), and descriptions are obtained. The observed peptides are screened by a user-configurable identification score threshold to avoid overestimation. Prior to the emPAI calculation, protein sequences from well-established databases were digested *in silico* and their sequences, MW, and predicted retention times were stored in an internal MySQL database. Compatible databases are the International Protein Index (IPI), UniProt, *Saccharomyces* Genome Database (SGD) and GenoBase; they can be easily expanded by adding new protein data using FASTA files. The observable peptides are obtained for each queried protein from the database using the extracted IDs. The peptides are further screened based on their MW and LC retention time ranges; these can be specified by the user to suit individual experimental conditions. The emPAI is calculated from the number of observed/observable peptides using the algorithm described above and presented as an HTML table.

ACKNOWLEDGEMENTS

We thank K. Komatsu, M. Matsui and N. Kono for technical advice and support. We also thank Drs L. M. de Godoy and M. Mann (Max-Planck Institute for Biochemistry, Martinsried, Germany) for access to the Mascot™ result files of their yeast LC-MS/MS data.

Funding: Yamagata Prefectural Government and Tsuruoka City.

Conflict of Interest: none declared.

REFERENCES

Corbin,R.W. *et al.* (2003) Toward a protein profile of Escherichia coli: comparison to its transcription profile. *Proc. Natl Acad. Sci. USA*, **100**, 9232–9237.

de Godoy,L.M. *et al.* (2006) Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biol.*, **7**, R50.

Ishihama,Y. *et al.* (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell Proteomics*, **4**, 1265–1272.

Lasonder,E. *et al.* (2002) Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. *Nature*, **419**, 537–542.

Liu,H. *et al*. (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.*, **76**, 4193–4201.

Lu,P. *et al.* (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.*, **25**, 117–124.

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. USA*, **77**, 1632–1636.

Petritis,K. *et al.* (2003) Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Anal. Chem.*, **75**, 1039–1048.

Rappsilber,J. *et al.* (2002) Large-scale proteomic analysis of the human spliceosome. *Genome Res.*, **12**, 1231–1245.

Shen,Y. *et al.* (2002) High-efficiency nanoscale liquid chromatography coupled on-line with mass spectrometry using nanoelectrospray ionization for proteomics. *Anal. Chem.*, **74**, 4235–4249.

Zybailov,B. *et al.* (2006) Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J. Proteome Res.*, **5**, 2339–2347.