RESEARCH ARTICLE

Staged-probability strategy of processing shotgun proteomic data to discover more functionally important proteins

Hong Xu¹, Guijun Ma¹, Qingqiao Tan¹, Qiang Zhou², Wen Su³, Rongxiu Li¹✉

¹ State Key Laboratory of Microbial Metabolism (Shanghai Jiao Tong University) and School of Life Sciences & Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

² Department of Bioinformatics and Biostatistics, College of Life Sciences & Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

³ Software Engineering Institute, East China Normal University, Shanghai 200062, China

✉ Correspondence: rxli@sjtu.edu.cn

Received September 30, 2011 Accepted November 22, 2011

ABSTRACT

Biologically important proteins related to membrane receptors, signal transduction, regulation, transcription, and translation are usually low in abundance and identified with low probability in mass spectroscopy (MS)-based analyses. Most valuable proteomics information on them were hitherto discarded due to the application of excessively strict data filtering for accurate identification. In this study, we present a staged-probability strategy for assessing proteomic data for potential functionally important protein clues. MS-based protein identifications from the second (L2) and third (L3) layers of the cascade affinity fractionation using the Trans-Proteomic Pipeline software were classified into three probability stages as 1.00–0.95, 0.95–0.50, and 0.50–0.20 according to their distinctive identification correctness rates (i.e. 100%–95%, 95%–50%, and 50%–20%, respectively). We found large data volumes and more functionally important proteins located at the previously unacceptable lower probability stages of 0.95–0.50 and 0.50–0.20 with acceptable correctness rate. More importantly, low probability proteins in L2 were verified to exist in L3. Together with some MS spectrogram examples, comparisons of protein identifications of L2 and L3 demonstrated that the staged-probability strategy could more adequately present both quantity and quality of proteomic information, especially for researches involving biomarker discovery and novel therapeutic target screening.

KEYWORDS lower minimum probability threshold, probability grade, TPP, low-abundant protein

INTRODUCTION

Liquid chromatography (LC) coupled with high-throughput mass spectroscopy (MS) is a powerful analysis tool for proteomic studies. The general approach to processing experimental data on peptides is to compare the MS spectrum with the theoretical spectra in specific databases (Yu et al., 2009) using search software based on different algorithms (Deutsch et al., 2008). Due to either the low quality of the spectra or the imperfection of the peptide-sequence database, the correctness of an assignment has taken precedence over the identification of the best match in a database. Great efforts have been exerted into achieving higher “correctness” and minimal “false” rate of peptide assignments in database searching (Patterson and Aebersold, 2003; Nesvizhskii and Aebersold, 2004). Usually, a “data filtering” step is applied to exclude the “incorrectly” assigned groups at the cost of losing the completeness of the proteomic information. Biologically important proteins related to membrane receptors, signal transduction, regulation, transcription, and translation are usually low in abundance and are usually identified with low probability. Most valuable proteomics information on them were hitherto discarded due to the application of excessively strict data filtering for accurate identification in MS-based analyses. In this study, the biological meaning of probability was rediscovered. We propose a novel staged-probability approach (SPA) for the

assessment of shotgun proteomic data in order to identify more functionally important proteins, exemplified by LC/MS datasets of rat liver proteins pretreated by cascade affinity chromatography and processed by the Trans-Proteomic Pipeline (TPP) software series (Tan et al., 2009).

RESULTS

Three probability grades of identified proteins

LC-MS and shotgun proteomic analysis strategy revealed that more total proteins could be identified with an increasing number of pre-fractionations separated before trypsin digestion. Using peptides with a probability (p) value greater than or equal to 95% as correct filtering conditions, the original LC/MS datasets of rat-liver proteomics by Tan et al. (Tan et al., 2009) identified 391 non-redundant proteins in unfractionated rat liver cytosol, 499 proteins in 2 fractions of the first layer affinity fractionation, 616 in 4 fractions of L2 fractionation, and 738 in 8 fractions of L3 fractionation.

Proteins identified at L2 and L3 were grouped using 16 intervals with minimum-probability values from 1.00 to 0.00. The number of identified proteins and the corresponding correctness rates of the assigned proteins were plotted against the minimum-probability intervals (Fig. 1). L2 and L3 datasets exhibited the same trend. The number of proteins identified was maximal at probability 1.00–0.99, dramatically decreased to a sixth of the maximal number at a probability

interval 0.95–0.90, and then increased slowly to a third of the maximal number at a probability interval 0.40–0.30. Correspondingly, correctness rate of identified proteins at each probability interval was more than 95% at a probability 1.00–0.95 and decreased to about 20% when the probability was from 0.95 to 0.20. Thus, the identified proteins could be graded into three stages with distinct correctness rates of 20%–50%, 50%–95%, and 95%–100% (Fig. 1; dashed lines). There is an apparent relationship between probability and identification correctness rates: a correctness rate of 95% corresponds to a probability of 0.95, a correctness rate of 50% corresponds to a probability of 0.50, and a correctness rate of 20% corresponds to a probability of 0.20. Thus, the biological meaning of probability is equivalent to the correctness rate value of protein identification at the corresponding probability interval.

The minimum probability of 0.95 is generally chosen as a critical point for filtering of protein data. In general, protein information identified with a probability of less than 0.95 has been regarded as unacceptable for publication by many proteome journals (Gasteiger et al., 2003). For L2 and L3 datasets, the total numbers of proteins in probability 0.95–0.50 and 0.50–0.20 were respectively 42% and 49% of the number of proteins with a correctness rate of 95%, thus accounting for nearly half of all identifications. Proteins identified with a correctness rate greater than 50% is equal to about 20% of the number of proteins with a correctness rate greater than 95%.

Figure 1. Number and protein identification correctness rate at every probability stage of rat liver proteomic data from layer 2 (L2) and layer 3 (L3) of the cascade affinity fractionation LC/MS and shotgun proteomic analysis.

Existence confirmation of low probability proteins

Since proteomic data of L2 and L3 were derived from the same original sample, proteins in the probability stage 1.00–0.95 with an identification correctness rate of more than 95% could be used for confirmation of the existence of proteins located at the lower probability of 0.95–0.20. We compared proteomic information in L2 and L3 at each probability interval and found that 20% of proteins located in L2 with a probability of 0.95–0.50 were among the proteins identified in L3 with a probability 1.00–0.95 and correctness rate greater than 95%. In addition, 27% of proteins in L2 at the 0.95–0.50 probability stage were also identified in L3 at a probability higher than their original probabilities in L2 (Table 1). Findings matched the predicted tendency (Fig. 1) that at least 4% of correct identifications, equivalent to 20% of correct identifications in probability stage 0.95–0.50 where 20% of total results are, may be lost when the probability threshold is set at 0.95.

Functional analysis and confirmation of low probability proteins

Setting probability to ≥ 0.95 with an identification correctness rate of 95% as the confirmation of protein existence, information from L2 and L3 can mutually prove the existence of functionally important proteins located at the low probability stage. By calculating the number of proteins in each functional type, we found that in both proteomic datasets, proteins relevant to signal transduction, transcription, and functionally unclear proteins at probability stages 0.95–0.50 and 0.50–0.20 comprised more than 50% of total identifications, whereas proteins relevant to regulation, translation, and cell adhesion corresponded to more than 40% of total identifications.

For example, signal-transduction proteins usually play a

significant role in triggering a biological response (Quon et al., 1996; Davis, 2000). If, they are membrane-related as well, the possibility of developing potential drug targets will be much larger (Slamon et al., 1987; Rabilloud, 2003). As shown in Fig. 2A, the amount of signal-transduction proteins in probability stages 0.95–0.50 and 0.50–0.20 were equal to that in probability stage 1.00–0.95. More importantly, a higher percentage of membrane-related signal-transduction proteins were observed in probability intervals below 0.95, especially in L3; in fact, the number and percentage of membrane-related proteins in 0.95–0.50 and 0.50–0.20 stages were about twice as much as that in 1.00–0.99. Information related to these proteins are too important biologically and functionally to be ignored. Thus, a study aiming to discover functionally important proteins, biomarkers, or novel therapeutic targets should be designed to pay attention to proteins identified with a probability below 0.95 as well.

By mutual confirmation of signal transduction proteins, 12% and 6% of signal transduction proteins with a lower probability in L2 and in L3, respectively, were confirmed in the other dataset. Three proteins in the probability interval 0.95–0.90 in L2 were identified in L3 at a probability greater than 97%, including Annexin A1 (involved in cell surface receptor-linked signaling pathway (Ashburner et al., 2000)), Brefeldin A-inhibited guanine nucleotide-exchange factor 2 (regulates ARF protein signal transduction (Charych et al., 2004)), and Ras-related protein Rap-1A (induces morphological reversion of a cell line transformed by a Ras oncogene (Kitayama et al., 1989)). Two proteins in the probability interval 0.80–0.70 and one in 0.30–0.20 of L3, which were all membrane-related proteins, were verified in L2 at a probability higher than 98%. They were Sorting nexin-1 (regulates degradation of epidermal growth factor receptor (Chin et al., 2001)), Adapter molecule crk (activator of Jun kinase activity (Abassi and Vuori, 2002)), and Probable G-protein coupled receptor 37. Findings indicate that signal-transduction

Table 1 Number and percentage of proteins identified in L3 that are present in lower probability stages in L2

Probability intervals	No. of proteins in L2	No. of L2 proteins found at probability ≥ 0.95 in L3	Percentage of L2 proteins found at probability ≥ 0.95 in L3	No. of L2 proteins with probability upgrade in L3	Percentage of L2 proteins with probability upgrade in L3
0.95–0.9	38	11	29%	0	29%
0.9–0.8	44	12	27%	2	32%
0.8–0.7	33	5	15%	3	24%
0.7–0.6	40	7	18%	4	28%
0.6–0.5	43	5	12%	4	21%
Average			20%		27%
0.5–0.4	47	4	9%	5	19%
0.4–0.3	73	2	3%	17	26%
0.3–0.2	114	4	4%	21	22%
Sum	432	50	12%	56	25%
Average			14%		25%

Figure 2. Numbers and percentage of functional proteins in L2 and L3. (A) Membrane-related proteins found among signal-transduction proteins in L2 and L3 datasets; and (B) percentage of probability upgraded proteins with $p \leq 0.95$ in L2 and ≥ 0.95 in L3.

proteins with probabilities lower than 0.95 were very diverse and of great importance.

Moreover, proteins in 13 functional types with a low probability were verified at an identification correctness rate $\geq 95\%$ or upgraded in probability. The proportions of upgraded proteins of different functional types are illustrated in Fig. 2B. More than 20% of low probability proteins in three types and more than 10% in other ten types were upgraded. Therefore, if the aim of a study is to discover proteins related to the cell cycle, protein processing, or structure, it cannot ignore the 25%–33% potentially correct information in the lower probability stages. For example, Glutathione S-transferase P was identified in L2 at 0.60–0.50 intervals and at 100% probability interval in L3. It was reported to be

present in the liver at very low levels (Okuda et al., 1987; Ashburner et al., 2000), and is known to act as an inhibitor of c-Jun N-terminal kinase (JNK1) signaling transduction by interacting with MAPK8, which functions in susceptibility to cancer and anticancer drug resistance (Wang et al., 2001). Since low-abundance proteins are always more fascinating and difficult to discover, any clue of their existence is worthy of attention.

Mass spectrum confirmation of low probability proteins

Annexin A1 is a signal transduction protein that was identified in the probability interval 0.95–0.90 of L2 and was confirmed in L3 with a probability of 100%. Figures 3A and 3B are mass

Figure 3. Mass spectrogram of Annexin A1 from Bioworks 3.3. (A) In L2 without data filter, (B) In L3 without data filter, and (C) in L3 with data filter $\Delta Cn \geq 0.100$; $Xcorr(\pm 1, 2, 3) = 1.90, 2.20, 3.75$; and #Top match = 1.

spectra of Annexin A1 in L2 and L3, respectively, as displayed by Bioworks 3.3 with no data filters. Even the basic filter setting of Xcorr or ΔCn would filter Annexin A1 out of the L2 proteomic results. Figure 3C is also an MS spectrogram of Annexin A1 in L3 with filters set as $\Delta Cn \geq 0.100$; Xcorr ($\pm 1, 2, 3$) = 1.90, 2.20, 3.75; and #Top match = 1. The differences in MS peaks are obvious.

Detailed information on peptides for identifying Annexin A1 is presented in Table 2. Fourteen peptides are likely to exist in sample L2, with a protein coverage of up to 40% (bold peptides). However, all of these were excluded by empirical basic data filters. When eliminating experimental interferences by cascade affinity fractionation, some peptides were accepted with a high probability in L3, yielding a protein coverage of 28% (underlined peptides). Further filtering accepted only one peptide (italicized peptide) as the correct one for protein confirmation. Therefore, about 93% of potential peptide information were discarded from L2 to yield L3 final filtered results.

DISCUSSION

In many functional proteome studies, proteomic data analysis is the first screening step in identification of biologically meaningful clues. By analyzing proteomic datasets with the staged-probability strategy, we can not only discover more proteins, especially more functionally important proteins, but also assess the correctness rate of potentially existing proteins. For biomarker discovery and novel therapeutic target screening projects, a 95% probability filter can be too strict because more than 20% of identified proteins with a correctness rate greater than 50% would be ignored. And as selected proteins have to be verified experimentally even if they were identified with a probability of 1.00 and zero theoretical error, a larger volume of MS proteomic data can be used for analysis. Obviously, our results indicate that it is worthwhile to assess proteomic information in lower probability stages for discovering any clue on potentially important proteins, instead of discarding them based solely on machinery scores.

Multiple factors may affect the mass spectrum and may lead to the omission of peptide and protein information in MS

testing, which could not be allowed if the proteins were potential functionally important targets. Since there is no single experimental method that it optimized for the proteins of interest, thorough and deep analysis of MS data by staged-probability may be a necessary supplement. Results of analysis for Annexin A1 in L2 and L3 by the Bioworks software also confirmed the correctness of the grading data produced by TPP.

The staged-probability strategy avoids hasty exclusion of valuable proteomic MS data based solely on computed probabilities. It also prevents misjudgments on experiments, since functional or low-abundance proteins cannot be discovered in higher probability data stages. By assessing low probability proteomic information, an untouched biological area may be unveiled. The proposed approach has great potential in the identification of biologically important proteins related to membrane receptors, signal transduction, regulation, transcription, and translation that are usually low in abundance.

MATERIALS AND METHODS

Experimental datasets

Original LC/MS datasets of rat-liver proteomics were collected using cascade affinity fractionation and Trans-Proteomic Pipeline (TPP) software (Tan et al., 2009). Briefly, the cascade affinity fractionation was composed of different biomimetic affinity ligand columns with a diverse combination of characteristics for affinity fractionation at cascade level. The original sample was loaded onto the first layer column of one affinity column (first layer of the system), yielding a flow-through and an elution fraction. These were separately loaded onto the second layer of two affinity columns (second layer of the system, L2 layer), which generated two flow-throughs and two elution fractions. Fractions were in turns separately loaded onto the third layer of four affinity columns (third layer of the system, L3 layer), producing eight fractions. MS results of all sample fractions from each one layer were combined. Tan et al. found that the number of proteins identified increased as more fractionation layers were used (Tan et al., 2009).

Datasets from L2 and L3 of the cascade affinity chromatography system were separately processed with SEQUEST (Eng et al., 1994) (v.28, Bioworks 3.3 software package, Thermo Electron) against the Rat International Protein Index (IPI) database (www.ebi.ac.uk; v3.73)

Table 2 Peptides of Annexin A1 detected in L2 and L3

	1–10	11–20	21–30	31–40	41–50
1	MAMVSEFLKQ	ACYIEKQEQE	YVQAVKSYKG	GPGSVSPYP	SFNPSSDVAA
51	LHKAIMVKGV	<u>DEATIIDILT</u>	<u>KRTNAQRQKI</u>	KAAYLQETGK	<u>PLDETLKKAL</u>
101	<u>TGHLEEVVLA</u>	<u>MLKTPAQFDA</u>	<u>DELRAAMKGL</u>	<u>GTDEDTLIEI</u>	<u>LTTRSNNQIR</u>
151	EITRVYREEL	KRDLAKDITS	<u>DTSGDFRNAL</u>	<u>LALAKGDRCE</u>	DMSVNQDLAD
201	TDARALYEAG	ERRKGTDVNV	FNILTTRSY	PHLRKVFNQY	RKYSQHDMNK
251	ALDLELKGDI	<u>EKCLTTIVKC</u>	<u>ATSTPAFFAE</u>	<u>KLYEAMKGAG</u>	<u>TRHKTIRIM</u>
301	VSRSEIDMNE	<u>IKVFYQKKYG</u>	<u>IPLCQAILDE</u>	<u>TKGDYEKILV</u>	ALCGGN

Bold: peptides identified in L2 (39.60%); Underlined: peptides identified in L3 without data filter (28.03%); Italicized: peptide accepted in L3 after filtering (3.76%).

(Kersey et al., 2004). Search results were subjected to statistical filtering using ISB/SPC Proteomics Tools-TPP V4.2. A correctness probability value equal to or higher than 0.05 was chosen as the basic filter condition for peptides.

Protein probability interval

The ProteinProphet module in TPP automatically graded the identified proteins using 16 intervals with minimum-probability values ranging from 1.00 to 0.00, together with the number of correctly and incorrectly identified proteins. Fourteen probability intervals were set based on each minimum-probability value. For example, the interval "1.00" contained proteins with a theoretically 100% probability of existence, while the interval "1.00–0.99" contained proteins with a probability of 99%–100% ($100\% > p \geq 99\%$). The numbers of correctly and incorrectly assigned proteins in every probability interval were summed, and the percentage of correctly identified proteins in each probability interval was calculated as the correctness rate. Both values were used to evaluate the quality and volume of protein information in association with the decreased probability.

Annotation

Functional annotation of all identified proteins was performed by considering information related to the terms "Function," "Keywords," and "Gene Ontology" from the UniProt Knowledge Base (UniProtKB) database (<http://www.uniprot.org>) (Ashburner et al., 2000; Carr et al., 2004; European Institute for Bioinformatics, 2005). To fully understand the proteins that exist at each probability interval, each protein was retrieved from the database and classified according to its function.

Comparison of mass spectrum in L2 and L3

To show the peptide information that were missed due to excessively strict data filters, the original mass spectra of a reclassified protein in L2 and L3 were determined using Bioworks 3.3. All corresponding peptide information were also gathered.

ACKNOWLEDGEMENTS

We acknowledge the financial support from the National S&T Major Projects of China (Key Innovative Drug Development, No. 2009ZX09306-008), National Basic Research Program of China (973 Program, Grant Nos. 2007CB936004 and 2009CB118906), the National Natural Science Foundation of China (Grant No. 30630012), Shanghai Leading Academic Discipline Project (No. B203) and Shanghai Science and Technology Innovation Action Program (Nos. 072312048 and 08DZ1204400). We also thank the Instrumental Analysis Center of Shanghai Jiao Tong University for SEQUEST searching. The authors declare no conflict of interest.

ABBREVIATIONS

LC, Liquid chromatography; TPP, Trans-Proteomic Pipeline (TPP) software series; UniProtKB, UniProt Knowledge Base

REFERENCES

Abassi, Y.A., and Vuori, K. (2002). Tyrosine 221 in Crk regulates

adhesion-dependent membrane localization of Crk and Rac and activation of Rac signaling. *EMBO J* 21, 4571–4582.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al., and the The Gene Ontology Consortium. (2000). Gene ontology: tool for the unification of biology. *Nat Genet* 25, 25–29.

Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. (2004). The Need for Guidelines in Publication of Peptide and Protein Identification Data. *Mol Cell Proteomics* 3, 531–533.

Charych, E.I., Yu, W., Miralles, C.P., Serwanski, D.R., Li, X., Rubio, M., and De Blas, A.L. (2004). The brefeldin A-inhibited GDP/GTP exchange factor 2, a protein involved in vesicular trafficking, interacts with the beta subunits of the GABA receptors. *J Neurochem* 90, 173–189.

Chin, L.S., Raynor, M.C., Wei, X., Chen, H.Q., and Li, L. (2001). Hrs interacts with sorting nexin 1 and regulates degradation of epidermal growth factor receptor. *J Biol Chem* 276, 7069–7078.

Davis, R.J. (2000). Signal transduction by the JNK group of MAP kinases. *Cell* 103, 239–252.

Deutsch, E., W., Lam, H., and Aebersold, R. (2008). Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol Genomics* 33, 18–25.

Eng, J.K., McCormack, A.L., and Yates Iii, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 5, 976–989.

European Institute for Bioinformatics (EBI). (2005). Available at: <http://www.ensembl.org>.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., and Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* 31, 3784–3788.

Kersey, P.J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004). The International Protein Index: an integrated database for proteomics experiments. *Proteomics* 4, 1985–1988.

Kitayama, H., Sugimoto, Y., Matsuzaki, T., Ikawa, Y., and Noda, M. (1989). A ras-related gene with transformation suppressor activity. *Cell* 56, 77–84.

Nesvizhskii, A.I., and Aebersold, R. (2004). Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discov Today* 9, 173–181.

Okuda, A., Sakai, M., and Muramatsu, M. (1987). The structure of the rat glutathione S-transferase P gene and related pseudogenes. *J Biol Chem* 262, 3858–3863.

Patterson, S.D., and Aebersold, R.H. (2003). Data analysis – the Achilles heel of proteomics. *Nat Genet* 33, 311–323.

Quon, K.C., Marczyński, G.T., and Shapiro, L. (1996). Cell cycle control by an essential bacterial two-component signal transduction protein. *Cell* 84, 83–93.

Rabilloud, T. (2003). Membrane proteins ride shotgun. *Nat Biotechnol* 21, 508–510.

Slamon, D.J., Clark, G.M., Wong, S.G., Levin, W.J., Ullrich, A., and McGuire, W.L. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 235, 177–182.

Tan, Q., Dong, D., and Li, R. (2009). A novel fractionation method prior to MS-based proteomics analysis using cascade biomimetic

- affinity chromatography. *J Chromatogr B Analyt Technol Biomed Life Sci* 877, 3799–3805.
- Wang, T., Arifoglu, P., Ronai, Z., and Tew, K. D. (2001). Glutathione S-transferase P1–1 (GSTP1–1) Inhibits c-Jun N-terminal Kinase (JNK1) Signaling through Interaction with the C Terminus. *J Biol Chem* 276, 20999–21003.
- Yu, K., Sabelli, A., DeKeukelaere, L., Park, R., Sindi, S., Gatsonis, C. A., and Salomon, A. (2009). Integrated platform for manual and high-throughput statistical validation of tandem mass spectra. *Proteomics* 9, 3115–3125.